A System to Assess the Achievement of Doctor of Pharmacy Students[1]

Nancy Winslade[2]
Hortensialaan 41, 3702 VE, Zeist, The Netherlands

PURPOSE

The purpose of this document is to make evidence-based recommendations regarding a system to assess outcome achievement of students enrolled in Doctor of Pharmacy (PharmD) programs in the United States. Data collected from the recommended assessment system could be used for the multiple purposes of program assessment and continuous improvement, pass/fail, or grading decisions (summative evaluation), and student feedback to maximize individual learning(formative evaluation). The emphasis of the system, however, is as part an institution's assessment plan that aims to continuously improve the quality of the educational programs offered at the institution(1). Particular importance is placed on developing a system that would be able to, with the American Association of Colleges of Pharmacy's (AACP) facilitation and upon request by a college or school, provide assessment data that could be used by individual institutions to compare achievement of their students with that of students from similar institutions. The recommended system is not meant to replace all student evaluation activities at a college or school, but is meant to provide a starting point for collection of assessment data for the primary purpose of quality assurance.

BACKGROUND

Over the past two decades, universities have faced an increasing demand to demonstrate the quality of the programs and services they provide to society(2-6). A number of models have been applied to quality assessment in higher education as reviewed by Madaus, Schriven, and Stufflebeam(7). The most widely accepted models consider a variety of factors when determining the quality of the university and its programs. For example, these models consider the quality of the entering students, faculty, curriculum, educational resources, and graduates. To facilitate the evaluation of these various factors, they are often categorized as inputs to the university program, processes, or the environment associated with the university program, and outputs generated by the university program(2,8). Figure 1 represents the application of this type of categorization to pharmacy degree programs.

It is important to note that this figure focuses on the **educational programs** of a college or school of pharmacy and that these programs represent only one of the three major mission components of most institutions. The other two responsibilities are generation of knowledge (*i.e.,* research) and service to society, and a comprehensive quality assurance system should evaluate the quality of all three components(3). The figure is consistent, however, with AACP's focus on facilitating the

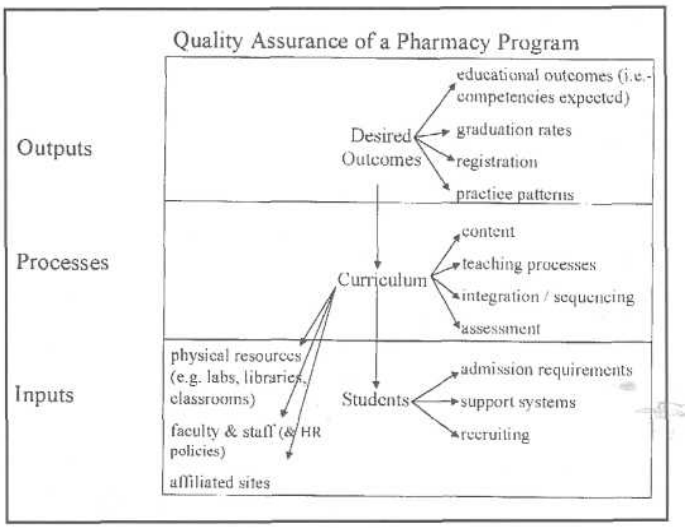

Fig. 1. Quality assurance system for pharmacy education programs.

establishment of processes to assure the quality and effectiveness of the **educational programs** designed, implemented, and monitored by colleges and schools of pharmacy(9). AACP also states that such processes should include local and national systems to assess the quality of both students and the program in general, thereby emphasizing that the former (*i.e.,* quality of students) is only one component of a quality assurance or program assessment system. This concept of the need for a comprehensive program evaluation system was reinforced at the 2000 AACP Institute and in the Guide for Doctor of Pharmacy Program Assessment 1).

The American Council on Pharmaceutical Education's (ACPE) accreditation standards also mandate that evaluation systems for pharmacy programs include assessments of program inputs, processes, and outcomes(10). Finally, the positions of both AACP and ACPE are consistent with the draft International Accreditation Standards for Basic Medical Education proposed by the World Federation on Medical Education(11). These include requirements for stating of expected educational outcomes, assessment of students, and quality assessment of the program.

Within this context of the requirement for a broad quality assurance program, focus has often been placed on the assess

---

[1]Paper prepared for the American Association of Colleges of Pharmacy.

[2]*Corresponding address:* Nancywinslade@compuserv.com

Am. J. Pharm. Educ., **65**, 363-392(2001).

ment of the quality of students(12,13). ACPE emphasizes this particular component of program evaluation by stating that "information regarding the effectiveness of the professional program in pharmacy, particularly in the form of student achievement, should be gathered systematically from sources such as students, alumni, state boards of pharmacy and other publics, professional staff of affiliated practice facilities, and a variety of other practitioners"(10). In practice, however, the greatest interest seems to lie in the evaluation of students' achievement as defined by assessment of their ability to meet desired educational outcomes(12,13). In pharmacy, this interest has resulted in the development of the AACP Center for the Advancement of Pharmaceutical Education(CAPE) *Educational Outcomes* suggested for of American pharmacy graduates(14), the AACP *Handbook on Outcomes Assessment*(15), and two reports from the AACP Council of Faculties on Teaching and Outcomes Assessment[3](16,17).

Despite the availability of these resources, a significant challenge remains regarding the assessment of the quality of students and their achievement. This challenge is the need to define quality through the development of standards against which student achievement can be compared on both an individual and class/school basis(5,18,19). To develop such standards, the level of expected performance in each of the desired educational outcomes should be defined through the development of relatively detailed performance indicators(2). Student performance can then be compared against these standards. However, developing such standards is a complex task as it is difficult to both define the realistic level of expected performance and then accurately and adequately describe this level of performance in a clear, concise manner(18,20).

An alternative, but complementary method for developing standards is through the use of benchmarking(21). Applied to student achievement assessment, this method identifies as the standards of performance the performances of students in the "best" colleges or schools of pharmacy. Individual colleges or schools can then compare the performance of their students to this standard of performance. The ideal situation is where student performance is compared both with the accepted national standards and with the performance of students at other similar colleges or schools of pharmacy. If information on performance throughout the program is obtained, then these comparisons can be used to identify points in the curriculum that should be focused upon during subsequent quality assurance activities. The value of comparing student achievement with standards and among schools is emphasized throughout this paper as a critical aspect of quality assurance. This emphasis is reflected in the focus on developing a student achievement assessment system that allows multiple colleges or schools of pharmacy to use common assessment formats and tools.

When developing a quality assurance system involving assessment of student achievement of outcomes, it must be recognized that significant changes have occurred in the formats and tools used to assess the achievement of students in the health professions. Key among these developments has been a move towards performance-based assessment, particularly when the resulting data is used to make summative decisions such as those required for graduation or licensure. Concomitant with this shift has been the development of the common belief that alternative formats of assessment, such as written tests, are less relevant to the evaluation of students' ability to fulfill desired educational outcomes (for example, see definition of performance assessment in references 9, 16, and 17).

These changes and perceptions have resulted in a large volume of research that addresses student achievement assessment. The majority of this literature comes from the field of medicine where the domains of research and assessment in medical education have been well established for decades. The resources accessible to the medical profession have led to the availability of the most complete and thorough investigations of assessment in the health professions at both the level of student evaluation and assessment for licensure/certification. Although professions such as dentistry, optometry, nursing, and pharmacy have had licensure-related requirements for assessment of practitioners for a number of years, the amount of peer-reviewed literature in the public domain that describes the rationale and psychometrics of these assessments is limited. Furthermore, where literature exists, it is clear that several of the health professions have relied heavily on the extensive literature from medicine when developing their assessment programs(22-25).

Relative to medicine, the majority of the other health professions are either in their infancy regarding the study of the development and assessment of expertise within their professions(26) or have focused on particular aspects of competency assessment rather than assessment of competency in its entirety. An example of the latter is the nursing literature's emphasis on teaching and assessment of critical thinking as the fundamental aspect of nursing competency(27-31). It is also important to note that the results of these initial psychometric investigations in the allied health professions, including pharmacy, tend to be consistent with the findings available in the literature published from the field of medicine(22-24,32).

This paper, therefore, relies heavily on assessment literature from the field of medicine with reference to literature from other health professions when specific differences, examples, or "best practices" offer unique information relevant to assessment of pharmacy students. This reliance on medical literature is not meant to minimize the importance or quality of the research completed by other health professions, but attempts to use the most complete, continuous literature available that relates to the assessment of pharmacy students.

Finally, although the paper does make substantial use of the literature from higher education, again the focus is more on literature from health professions education. This recognizes that the goals of education in the health professions are weighted differently than the goals of general, higher education in that career preparation assumes a greater emphasis in the health professions(3,33). This, in turn, leads to an emphasis on professional educational outcomes that are based on competencies expected of the health professional relative to the general educational outcomes expected of university graduates and educated citizens. Given that the educational outcomes defined as required of pharmacy graduates by AACP(14) follows this emphasis on professional outcomes, literature that focuses on the assessment of these outcomes has been used to a greater extent than the literature from general higher education.

PRINCIPLES AND TERMINOLOGY

Several steps must be followed when developing a student achievement assessment system, beginning with defining the

[3]For the remainder of the document, desired outcomes will refer to the AACP CAPE *Educational Outcomes*(14), and student achievement assessment and student outcome assessment will be used interchangeably.

desired educational outcomes required of graduates, including the setting of standards through the detailing of the levels of expectation and contexts within which graduates should be competent(13,19,34-37). AACP has begun this process of standard setting by defining the professional practice-based and general ability-based educational outcomes desired of graduates, including very brief definitions of the levels of expectation associated with the general ability-based outcomes(14,38). Once these contexts and levels have been specified in sufficient detail, the next step is the development of an assessment blueprint that requires decisions regarding both the weighting of the various outcomes in the overall assessment system and the selection of the types of formats that will be used for assessment of each outcome. The first step should be based on either actual or desired professional practice patterns depending on the degree to which the educational programs are attempting to guide change in professional practice. For the latter step, a choice must be made among assessment formats such as written assessments (*e.g.,* multiple-choice, written essays, short answer questions), oral examinations, in- training assessments, or demonstration projects.

CRITERIA FOR SELECTING ASSESSMENT FORMATS

When selecting the most appropriate formats to be included in an assessment, five criteria must be considered(39). These are the validity, reliability, educational impact, feasibility, and acceptability of the assessment format. In addition, since the primary goal of the system being suggested in this paper is to provide data that can be used for quality assurance, an additional selection criterion must be the potential usefulness of the assessment format by multiple colleges and schools of pharmacy. This latter point was addressed in the Background section. The following very briefly addresses each of the first five criteria.

1. Validity

Most simply stated, validity considers how well the assessment format measures what it proposes to measure. For example, it must be determined whether a rating form that a preceptor uses to assess a student's ability to provide pharmaceutical care really measures this outcome or, in fact, measures the preceptor's assessment of the student's personality, communication skills, or other abilities. There are two general approaches to validity: an indirect approach that looks at evaluating the validity of students' scores on a specific assessment by examining if the patterns of results are consistent with expectations (*e.g.,* do expert practitioners score higher than final year students who, in turn, score higher than first year students)(40), and a second, direct approach, that focuses on ensuring that valid results are obtained through careful selection, development, and design of the assessment format and tool(41,42).

Indirect methods are more frequently used than direct measures, and the most commonly used indirect measure is the correlational analysis that examines criterion validity. In this method, students' results on one assessment are compared with their performance on another assessment that theoretically measures the same outcome. The problems encountered with this type of validation are twofold. First, it presumes that there is a gold standard assessment format that definitively measures the relevant outcome. Second, the results of such comparisons usually yield intermediate correlations: implying either that one or both of the two tools was poorly designed, or that they both may be measured similar(but perhaps different) constructs(as an example from pharmacy see reference(43).

In the former situation, the use of tools that are poorly designed can lead to very unreliable results with the performance of students scattered in an inconsistent manner. Obviously, it is difficult to obtain strong correlations between two measures that lead to such scattered results. For this reason, reliability calculations should be included in studies that attempt to correlate scores on different tools and correlations should be corrected for unreliability(44). Without such information, interpretation of correlation coefficients, especially low correlation coefficients, is extremely difficult. For example, consider the situation where a faculty member develops two evaluations to assess students' ability to manage ethical dilemmas.

The first format is a written exam that presents a series of scenarios and students are asked to identify ethical principles involved in the scenario, such as beneficence and confidentiality. The second format involves role playing with fellow students where the faculty member uses a rating form to assess the students' competency at managing the dilemma. When students' scores on the two evaluations are compared, the correlation coefficient (uncorrected for unreliability) is 0.45 ($P< 0.05$). This result would be very difficult to interpret since it is not known which of the two tools is really the best measure of the desired outcome and the intermediate correlation could mean that they both are measuring (rather poorly) the same construct or they are measuring different, but somewhat related constructs. From this example, it should become clear that these types of results lead to an inability to draw any definite conclusions about the validity of the assessment tool.

These problems have resulted in a greater emphasis on direct validation methods where critically evaluated literature on both the theory of the outcome being assessed and the methods to assess the outcome are used to guide decisions about the assessment format and item type selected (referred to as construct validity). For example, recent theories on how medical expertise evolves have been used to develop new types of assessment formats that are both consistent with these theories and lead to improved psychometrics (see later discussion on key-features testing). A second component to direct validation methods is the use of detailed blueprints to ensure an appropriate and balanced sampling of all the outcomes required of students (referred to as content validity)(41,44,45). Although controversy still exists regarding the best way to document the validity of an assessment format and tool, evidence of validity is considered critical when selecting formats to be included in a student assessment system.

2. Reliability

Reliability refers to the reproducibility of a student's results obtained with a specific assessment tool and addresses the question: how consistent would a student's score be if (s)he completed this assessment multiple times under different situations, or with different assessors, or with equivalent (but slightly different) questions? In other words, how much confidence is there that the results obtained are generalizable to other assessment situations? Historically, a strong focus has been placed on one component of reliability: objectivity, which is the consistency of scores assigned by two or more raters/graders. This focus has resulted in a number of problems, including a confusion about the relationship between

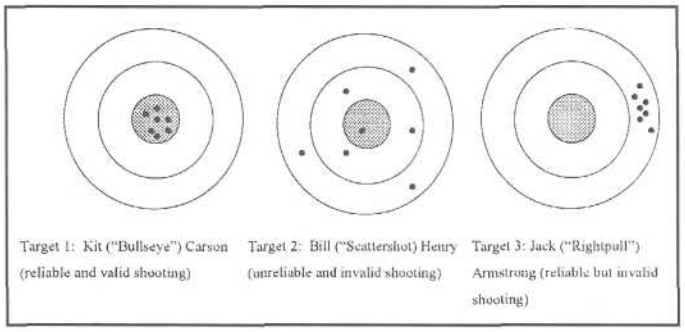

Fig. 2. The relationship between validity and reliability(35) (Reprinted with permission from Prentice Hall).

reliability and objectivity with, on occasion, the terms being used interchangeably as if no other factors contribute to the reliability of a testing format(46).

To avoid this confusion, the term generalizability is frequently used to refer to the global reliability of a testing format(45-47). Regarding the relationship between global reliability and objectivity, multiple reports clearly indicate that objectivity is not the primary determinant of the generalizability of an assessment format(48,49) and that, in fact, inter-rater consistency is one of the factors that can be most readily managed in many assessment formats(39,46). Therefore, although it is important to consider the inter-rater reliability of an assessment format, consideration must be given to the other factors that contribute to the global reliability or generalizability of the format. For example, the variability in student performance seen with formats that include too few assessment questions has been documented to have a much greater impact on global reliability, and be much more difficult to control, than inter-rater reliability(50,51).

Another confusing relationship is between validity and global reliability. Linn and Gronlund(35) provide a simple figure that explains this relationship: it is possible for an assessment format to be invalid, yet still reliable, but an assessment format can never be a valid, yet unreliable measure of a particular outcome (Figure 2).

3. Educational Impact

Multiple researchers have documented the remarkable steering effect that the assessment system has on student learning: students learn what they will be tested on and do not learn on what they will **not** be tested(52-54). This belief is best stated by Van der Vleuten, *et al.* as: "In educational practice we tend to ignore a very strong and lawful relationship between student assessment and student learning. The lawful relationship is that assessment drives learning.…..students will do whatever the examination programme tells them to do and they will not do whatever the examination programme does not reward. For the students, the examination programme *is* the curriculum."(55) This lawful relationship means that the assessment system used in a school must be consistent with the desired educational outcomes: if there is a conflict between the two then the assessment system will dominate and direct the real learning of students. Van Berkel(56), Norman(53,57), and Van der Vleuten, *et* al.(58) have discussed the assessment systems that are most consistent with the characteristics desired of graduates of health care professional programs such as a sound, well-integrated knowledge base, self-directed learning abilities, clinical reasoning skills, and communication skills.

The most important principle gained from this literature is that, in order for students to acquire the above characteristics, the assessment system used should encourage deep, integrated learning. Examples can be taken for each of the components of an assessment system, such as assessment regulations, scheduling, content, and formats. For assessment regulations, if "grades" in the individual biomedical sciences are much more heavily weighted than "grades" in the integrated courses, then students will focus on learning in a subject-oriented rather than integrated manner, *e.g.,* the hidden curriculum(59). If exams are scheduled in a separate content versus cumulative content manner (*e.g.,* the midterm examines the first half of the content while the final examines only the last half), then students may focus on shorter term learning and retention(60,61). The same is true if exams are scheduled to all occur within one short time frame. In this situation, students will tend to surface learn for short-term retention just before taking each exam. Regarding assessment content and formats, the most simple lesson is that the assessment system should focus on content, formats, and tasks that require students to integrate and apply their learning rather than simply recognize detailed facts that can easily be memorized and forgotten(37).

4. and 5. Feasibility and Acceptability

Although it is well accepted that the first three criteria discussed above are important to consider when developing a student assessment system, it must be emphasized that implementation of such centralized systems requires substantial resources, commitment, and change by the faculty of a college of school. If these factors are to be addressed appropriately, then two more criteria must be considered when selecting an assessment format. These are the feasibility and acceptability of the format to both faculty and students. From primarily the faculty perspective, there is no value in suggesting a system that requires the development and implementation of a four times per year objective structured clinical examination (OSCE) to a college or school with 20 FTE and no access to expertise in testing. Nor is it appropriate to recommend that rotation preceptors complete a five-page checklist on a daily basis for assessment of each of their students. Efficiency is key to the assessment format, the tools associated with the format, and the use of the data collected via the assessment. Centralization of assessment and the development and use of common formats and tools would maximize such efficiencies, and hopefully the faculty's perception of the feasibility and acceptability of the assessment system.

Student acceptance of the assessment system also depends partially on the efficiency of the system. The system must contribute to student learning and not detract from it by requiring students to dedicate excessive time to assessment activities(62). Perhaps more important, however, is that the assessment format, content, and tasks appear relevant, realistic, and fair to the students(35,37,63).

Summary: *When selecting the assessment formats to be included in a student assessment system that focuses on providing data for quality assurance, decisions must be based on the literature documenting the validity, global reliability, impact on education, feasibility, and acceptability of the formats and the potential for use by multiple colleges or schools.*

COMPETENCE VERSUS PERFORMANCE

When selecting assessment formats or tools, a presumption is often made that assessments that require students to demonstrate an ability or skill are better able to predict future, real-life

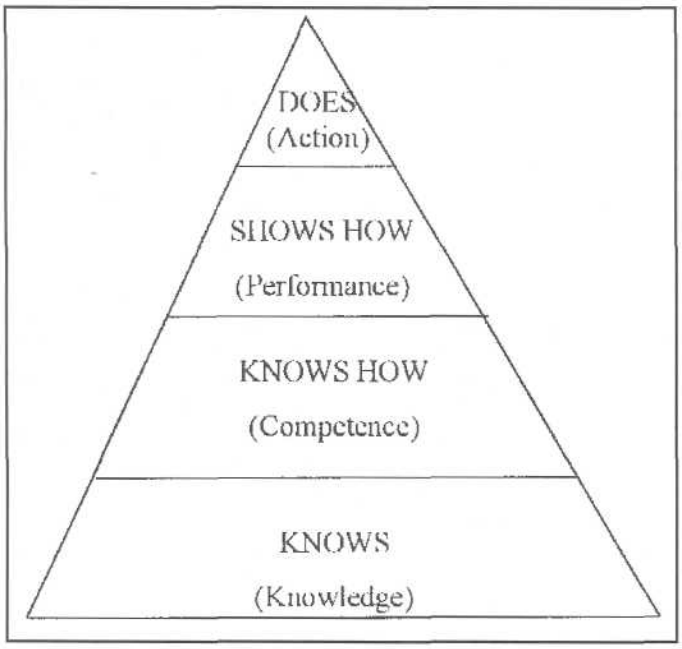

Fig. 3. Miller's pyramid(64) (Reprinted with permission from Academic Medicine and the Association of American Medical Colleges).

performance than other measures such as written assessments (*e.g.*, they have better predictive validity). This presumption is based on the belief that such assessments are more authentic and capable of testing complex skills, and that the ability of a student to *show how*(s)he can do an activity is predictive of his/her likelihood of doing this activity in real practice. In higher education, and in early work in medicine, these assumptions resulted in demonstration-based assessments being defined as performance-based assessments(37,64). In medicine, Miller's(64) pyramid framework for clinical assessment of students provided the basis for these definitions. AACP has adopted such definitions in much of its work(1,9,15-17). In more recent literature in the health professions, however, and particularly in the field of medicine, a clear distinction is being made between assessments of student **competence** relative to assessments of student/practitioner **performance**(53,65-67).

These definitions in medicine have evolved from Miller's(64) original definitions of competence as *knowing how* to do something and performance being able to *show how* to do something (Figure 3). In current medical assessment literature, competence is defined as what students are able to do in a maximum effort, testing, non-real-life environment, while performance is what students or practicing physicians do repeatedly in real-life practice with real patients: in other words, their average way of working on a regular basis(51,67,68). One reason for this distinction relates to the ability of assessments of students to accurately predict future, real-life performance of graduates.

Although it has been traditionally believed that lack of predictive capability of an assessment format meant that the format was not a useful or valid measure, in reality there may be many factors in real-life practice that prevent or inhibit a professional's ability to perform to his/her maximal ability, in the way that (s)he considers to be ideal, or in the way the (s)he performs during a testing or educational environment(65). For example, management might have a different philosophy and require the professional to alter his/her practice style to fit this philosophy, human resources may be insufficient to allow time to perform in the way that the professional desires, or payment systems may not reward ideal practice behaviors.

All of these factors make it unrealistic to expect perfect predictive correlations between measures of competence and measures of performance, as defined by medicine(65). Given these considerations, the question arises as to whether health professional students can ever be assessed on their true performance or whether they always function within an artificial, or competence-based environment. Newble suggests that, to the point of graduation, medical students are primarily functioning in an artificial environment and, therefore, assessment primarily concerns competence and not performance(69). Others in the field of assessment of medical students argue that during extended clinical rotations students adopt behaviors and attitudes that are truly reflective of their natural or real-life behaviors (personal communications, L. Schuwirth, MD, PhD, August 2000).

This argument would support the belief that assessments of routine performance during rotations do **not** represent an artificial, testing environment and more closely reflect real-life performance. However, in pharmacy it must be recalled that experiential rotations are relatively short in length and most often are offered in sites that represent ideal practice situations with trained preceptors dedicated to the advancement of pharmacy practice. In these sites, students often perform functions that represent ideal, advanced level practice such as extensive patient assessments, teaching of allied health care professionals, and development of protocols or proposals. These sites clearly differ from the average practice site in which licensed pharmacists may be expected to function. Therefore, although performance of students on experiential rotations may be the best predictor of future, real-life performance, it is unreasonable to expect perfect correlations between measures of performance on rotations and performance in real-life practice.

For all of the aforementioned reasons, medicine, as the profession with the most expertise in competence and performance assessment, is careful to distinguish between assessment formats that measure competence within artificial environments and those that measure performance in real-life situations. This leads to some confusion as, according to current definitions, some assessment formats that have been traditionally termed performance-based (*e.g.*, simulated patients administered via OSCEs) are really measures of competence rather than performance. The important issue, however, is not terminology and whether an assessment format is labeled as performance- or competence-based. The key issue is determining which assessment formats best predict actual performance of health practitioners in real life. This issue will be addressed later in this document when discussing the advantages and disadvantages of the different assessment formats.

Summary: *When selecting the assessment formats to be included in a student achievement system, decisions should consider the literature that examines the usefulness of the format as a predictor of future performance in real-life practice. This literature can either be theory-based (*e.g.*, direct validity) or research-based (*e.g.*, indirect validity and correlational studies). It should not be presumed that demonstration-type assessments are superior to written assessment formats.*

ASSESSMENT FORMATS VERSUS ASSESSMENT TASKS

When reviewing the usefulness of various assessment formats such as written examinations, portfolios, or simulations, several authors have attempted to categorize assessment formats

according to Miller's(64) levels of competence (see Figure 3)(68,70). These authors often attempt to fill in the four layers of Miller's pyramid with the assessment formats that they believe most appropriately measure a student's ability to *know, know how, show how,* or *do* a particular outcome(70). A difficulty arises when trying to create such categorizations, however, in that these categorizations are sometimes based on perceptions and the manner in which certain formats of questions are **most commonly** used, and not necessarily on current literature that documents the usefulness of the various formats. For example, there is a general belief that selection-type written assessments, and in particular multiple-choice question (MCQ) formats, can only be used to assess a student's ability to recognize information (see definition of performance assessment in reference 9). In other words, some believe multiple-choice questions can only be used to assess the bottom layer of Miller's pyramid: *knowing*(70).

However, this assumption relates to the manner in which multiple-choice questions are often written. They ask students questions that require recognition of a fact as opposed to focusing on understanding, integration, or application of information. That specific questions are often written in one particular way does not mean that the format is not useful for assessing higher levels such as *knowing how* within Miller's pyramid. On the contrary, current literature in assessment of health professionals is quite clear that it is **not** primarily the format of a question that determines the level of complexity being assessed. Instead, it is the specific task required of the students that makes this determination(44,71,72).

For example, an oral examination could require students to integrate and apply knowledge from multiple subjects to identify a paper patient's drug-related problems, or it could require students to state the two most common bacteria that cause community-acquired pneumonia. Therefore, in this latter question, although the format of the assessment is an oral examination (which is often presumed to be a superior assessment format than, for example, multiple-choice questions), the specific task requested of the student requires only recall of factual information. By contrast, recent developments in the field of multiple-choice question testing have modified the presentation and task requirements of these types of questions to ensure that they are assessing more complex competencies such as integration, knowledge application, and decision making(44,73-75).

These developments are discussed in detail in the sections below on written assessments. Despite this argument, it must be recognized that certain limitations do exist for certain assessment formats. For example, advanced-level standards for oral communication skills can not be assessed via objective, written assessment formats (*e.g.,* true/false or multiple-choice questions) regardless of how well these questions are written(67).

Summary: *When selecting the assessment formats to be included in a student achievement system, a key factor to consider is that it is not only the assessment format that determines the level of outcome/competency that is being assessed. This level is determined primarily by the specific task required of the student by the individual questions within the assessment format. Again, assumptions about the superiority or inferiority of specific assessment formats should be thoroughly investigated.*

ASSESSMENT ADMINISTRATION METHODS

Objective Structured Clinical Examinations (OSCEs)

Of the recent developments in competency assessment, one that is frequently discussed is the use of OSCEs. Developed in the 1970s, the OSCE was designed to overcome certain challenges faced by examiners, such as lack of standardization of patients presented or questions posed to students, insufficient numbers of tasks required of students and variability in grading of students(76). Therefore, although OSCEs are often labeled as an assessment format, they are really a test administration procedure that maximizes the psychometric characteristics of the assessment data collected(77). Any number of specific formats and tools can be incorporated into such a test administration system including performance of procedures on models(58,78), completion of written multiple-choice questions(79,80) or oral examination questions, or the demonstration of more complex, integrated skills during the management of problems presented by standardized, simulated patients(24,32,55,79,80). When discussing such station-based examinations, confusion often arises about the relationship between OSCEs and the use of simulated, standardized patients with some equating the two as if OSCEs always and only make use of standardized patients. In medicine, and based on the name of objective structured **clinical** examinations, the focus on OSCE-administered exams is usually on assessment of clinical skills and provision of medical care.

To provide students with the most realistic assessment setting, such examinations frequently use standardized patients as the format to assess students' competence in providing this medical care(77). According to the outcomes being assessed and the exam blueprint, in most situations one standardized patient presents in one station with that station assigned to focus on the assessment of the students' competency in several desired outcomes(77). At least one national examining body, however, combines the use of such standardized patients with other assessment formats in their OSCEs via the use of couplet stations(80). In this format, students perform specific tasks with a standardized patient in one station and then answer multiple-choice questions or short answer questions about the patient in the "coupled" station. In this situation, two assessment formats are used within the OSCE administration system.

It should also be remembered that standardized patients can be used for assessment within other administrative systems. For example, standardized patients can be sent into doctor's offices(81,82) or pharmacies (personal communications, Dr. L. Muzzin, August 1993) to assess the actual, real-life performance of health professionals, or they can be used to develop students' clinical/psychomotor skills and to provide formative feedback(78,83-85). Such examples help clarify the difference between the OSCE, which is an examination administration system, and standardized patients, which are used as a specific assessment format.

When considering the criteria for any assessment format recommended for inclusion in an assessment system, it is not appropriate to apply most of these criteria to an administration system (rather than a specific format) such as an OSCE. This is because the validity and global reliability, and to a certain extent the impact on student learning, acceptability, and feasibility, depend on the specific assessment format and the task required of the student in the format. Therefore, it is not possible to make a general statement that an OSCE is a valid, reliable, or superior way to assess students' achievements. This would imply that it is irrelevant whether the OSCE is com

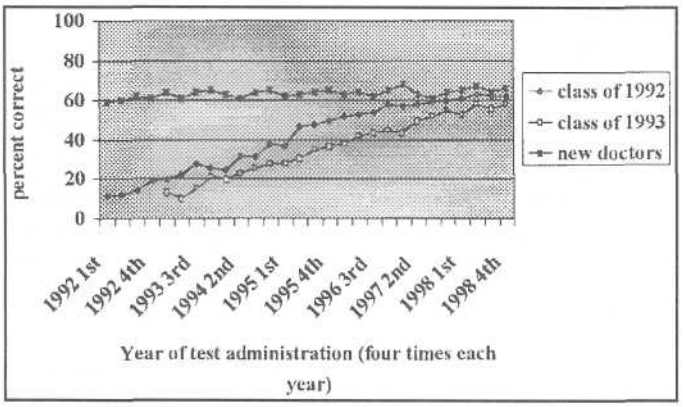

Fig. 4. Example of typical manner of reporting results from progress testing.

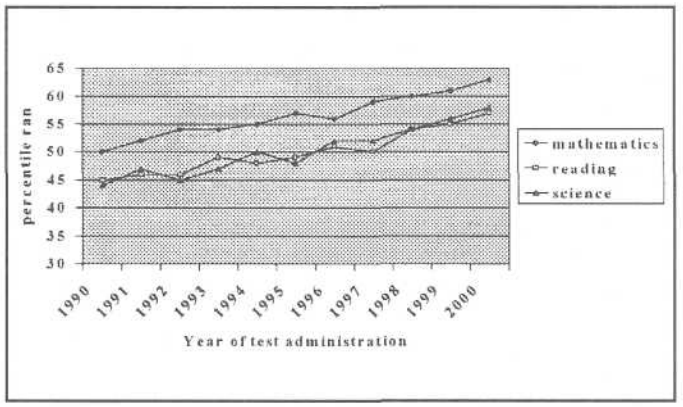

Fig. 5. Example of typical format for presenting results from standardized testing.

posed of two stations using one standardized patient and one multiple-choice question, or whether the OSCE uses 20 stations each of which requires students to perform a specific skill on a different standardized patient.

The summarizing of the literature on OSCEs is complicated by this fact, and important issues such as how many stations should be included and how long these stations should be depend on what is being assessed, the assessment formats, and the tasks required of students within the OSCE stations. For example, even if all stations use standardized patients, the time allotted for each station would be dependent on whether the student is required to complete a detailed history and physical to arrive at a diagnosis, or whether the student must complete a short assessment of the patient's cardiovascular system. Given this variability, only general statements can be made such as that, relative to unstructured assessment of students' demonstrations of skills, an OSCE administration method tends to increase the generalizability of the assessment data collected as students are tested on a greater number and range of scenarios and the assessment criteria are standardized. Detailed recommendations regarding the use of an OSCE administration method are discussed in the sections reviewing standardized patients.

Progress Testing

A second important development in assessment administration systems is that of progress testing(56,60,61,86-88). In this system, multiple equivalent assessments are developed based on the **final level** of functional knowledge, understanding, and knowledge application expected of students **at the time of graduation.** These equivalent assessments are then administered at regular intervals throughout the entire duration of the curriculum. For example, Maastricht Medical and Health Sciences Schools and the School of Medicine at the University of Missouri Kansas City administer a different, but equivalent written test of relevant knowledge four times per year to all students in the non-clinical years of their programs(61,86,87). Figure 4 shows the typical manner of reporting results of such testing. Since it is the **final level** of knowledge that is tested, students in the early years obviously score poorly on such assessments but scores increase steadily over the curriculum as students increase their functional knowledge base(58,61). With this type of assessment administration, detailed formative feedback on a single test or trends in performance can be provided to students to guide their subsequent learning. Cumulative year-to-year results and trends can also be used for the purpos es of quality improvement(58,61) and, if multiple colleges use similar assessments, comparisons can be made among these colleges(58,61).

The theory upon which progress testing is based aims to minimize the negative steering effect that the assessment system can have on student learning (see above discussion). If students are required to take exams that cover the entire range of knowledge required of graduates, and the content on each exam is dissociated from the specific content that the students are learning at a given point in the curriculum, then specific, strategic studying to the exam becomes a difficult and unproductive task(61). Such an administration system, therefore, encourages students to adopt deep learning methods that facilitate understanding and retention of knowledge and skills(60). At present, progress testing is used in a number of both traditional and problem-based health professions schools in the United States and the Netherlands(56,60,61,87-89). This use in both types of programs reflects the importance of outcomes requiring deep and self-directed learning among both traditional and problem-based health professional programs.

It is important to realize that this form of progress testing differs from the standardized testing frequently used in the United States for kindergarten through Grade 12(5). In standardized testing, standards of minimal performance and specific tests are set and developed for each grade. Each test, however, assesses different content and performance as defined by the standards. On a yearly or twice yearly basis, each grade level is assessed with the specific test for that grade and, most frequently, the same test is used year after year for each grade (*i.e.,* the Grade 3 students in 1999 take the same test as the Grade 3 students tool in 1998 and as the Grade 3 students in 2000 will take). Scores are calculated as the percentage of students in each grade that meet the expected level of minimal performance. Graphed results, therefore, typically indicate changes in performance of different groups of students (Figure 5). This differs from progress testing where each line represents an increase in knowledge of one group of students (or one individual student) over time. Perhaps the best analogy to draw for progress testing results are the normal growth curves used for monitoring the development of infants and toddlers. On these graphs, the height or weight of an individual child is plotted at various points during the child's development and the values compared with the normal values expected to ensure that the individual child is developing as expected. Similar to progress testing, such growth charts can be used for individual children or for groups of children to compare, for example, the normal growth patterns of Dutch relative to Spanish children.

By analogy, for quality assurance purposes, class performance on progress tests can be compared within schools or among schools if similar assessments are used at a number of schools.

Returning to the comparison of progress testing with standardized testing, it is important to realize the differences in the way results are recorded and reported. They are important as, in standardized testing(Figure 5), the performance of the reference group appears to improve from year to year (for example, each year Grade 3 students perform better and better) while for progress testing the performance of the reference group (*e.g.,* third year students) remains relatively constant over a number of years (Figure 4). Although supporters of standardized testing would suggest that the improvement in performance from year to year represents a real increase in performance that has resulted from improvements in teaching or curriculum, another explanation exists(5). When such results are used for accountability and possibly funding purposes, teachers can narrow their teaching to focus on preparing students to perform well on the standardized tests(5). This can result in a limitation of the learning of students. Although this teaching-to-the-test can be positive if the standards are well written and reflect ideal learning, teaching-to-the-test is not considered an approach that maximizes student learning(5). With progress testing, however, this same increase in performance for a specific reference group is not seen, indicating that there is neither a real increase in performance nor a teaching-to-the-test effect. This difference is understandable since the tests administered are equivalent but not identical and the level of content assessed represents a final, integrated level rather than the level taught by any individual teacher. Similar to student studying to the test, teaching-to-the-test, therefore, is a difficult and ineffective process.

Since progress testing is an administration method, then similar to the situation with OSCEs, it is not correct to make a general statement regarding the psychometric characteristics of progress tests. These types of analyses can only be completed on progress tests that specify the assessment format and the task required of the student in the format. For progress testing, the literature to date documents the validity, generalizability, feasibility, and general acceptability of using either true/false or multiple-choice question-based formats of assessment to evaluate the progressive increase in functional knowledge of medical and health sciences students over the duration of the curriculum(56,60,61,86-88). Regarding impact on student learning, evidence is also available that students follow different, more desirable, learning strategies for tests administered in a progress form as compared to more traditional forms where test content is directly linked to what has just been learned in the curriculum(60).

What is not available is literature examining these characteristics for progress testing in the demonstration-type formats such as repeated assessments in simulated environments or with standardized patients. For several reasons, the usefulness of progress testing for these formats is questionable. First, the skills that are assessed in such formats are often taught in a part-task training format(90). This means that, for example, physical examination skills are taught in sections where students learn to first do a head and neck exam, and then move on to chest, then cardiovascular examinations, etc. Given that the progress test would assess competency of performance of the whole task of physical assessment, the majority of the test (and the extensive time, training, and resources required) would be wasted on earlier students who would only be guessing at correct procedures. These assessments, therefore, would offer students the repeated opportunity to practice skills incorrectly during the progress tests. This, in turn, could result in the reinforcement of improper techniques that require unlearning at later points in the curriculum. Second, even if progress testing was appropriate for some skills that are taught in a whole task manner, again the resources required to run multiple, demonstration-type assessments over the course of the year are beyond those available to most schools. Such an administration method for demonstration-type assessments is, therefore, not feasible.

Summary: *When developing an assessment system, it must be recognized that OSCEs and progress testing are administration methods and not assessment formats. As such,* ***specific*** *statements regarding the desirability or superiority of OSCEs or progress testing can not be supported by empirical evidence. Recommendations regarding OSCEs and progress testing should only be made when the assessment format and task to be evaluated in the OSCE*(*e.g., using simulated patients to assess clinical skills) or progress test* (*e.g., using multiple-choice question-based written tests to assess functional knowledge) are specified.*

REVIEW OF ASSESSMENT FORMATS

> "It seems important to start with the forthright acknowledgement that no single assessment method can provide all the data required for judgment of anything so complex as the delivery of professional services."
>
> *G. Miller, 1990.*

Miller's comments during his address at the Research in Medical Education conference are as relevant today to the assessment of health profession students as they were in 1990(64). Despite many advances in the formats and tools used to assess students' achievement and competency, no single, perfect format has been developed that can assess the range of outcomes required of students in the health professions. When reviewing the different formats to identify those that should be included in an assessment system for pharmacy students, the challenges discussed in the section on assessment formats versus tasks should be remembered. In view of this discussion, rather than discussing each level of competence and which formats best "fit" into Miller's pyramid, it is more appropriate to review each of the formats in relationship to their usefulness as assessments of the various levels of competence/performance(64).

WRITTEN ASSESSMENTS

Knowledge and Understanding

Written assessments are usually divided into a number of categories such as objective tests (matching, fill in the blank, true/false, multiple-choice question), short/restricted answer questions, or essay questions(35). It is widely accepted that the most valid, reliable, feasible, and acceptable way to assess the knowledge base of students is through the use of well written multiple-choice question examinations(35,37,64,86,91,92). Two aspects of this statement must be analyzed when considering such a format for inclusion in an achievement assessment system for pharmacy students. The first is whether it is important to assess a student's knowledge base and the second is whether multiple-choice questions are really the best format to

use to assess this knowledge base. The first point relates to construct validity and, from the direct validity perspective, there is substantial literature that supports the central role of knowledge in competence(26,93,94). Modern theories in cognitive psychology and research from a number of fields have examined the development of expertise and have demonstrated that the extent, structure, and use of knowledge change as practitioners gain experience(94,95). From the field of medicine, it has been shown that experts have a largely experience-based knowledge that contains less detailed biomedical (*e.g.*, pathophysiology) knowledge and more contextual, functional knowledge(93,94).

This contextual knowledge is developed in phases beginning at a novice level where detailed biomedical knowledge dominates. The second stage represents additional understanding and application of knowledge as students gain initial experience with patients and compile their detailed knowledge into causal models and explanations for disease and treatment. With additional patient experience, these causal models lose biomedical detail and focus on illness scripts that incorporate more functional knowledge and are consistent with typical and atypical disease presentations and treatments. Finally, specific memorable patient encounters are incorporated into these illness scripts and experts' management of patients relies mainly on rapid recognition of similarities of new patient presentations with previous encountered presentations. Experts are still able to "unfold" their knowledge, however, to recall more detailed biomedical knowledge when faced, for example, with complicated patients or those who do not readily match with existing illness scripts(94). Extensive, well-connected, experience-based knowledge, therefore, remains a key requirement for medical expertise.

Substantial indirect validity evidence also exists to support the importance of knowledge. A number of studies have been completed that report intermediate to good correlations between student performance on knowledge tests and their performance on alternative assessments designed to evaluate clinical competence such as patient management problems(96), extended long answer questions(50), performance of clinical skills(97), and simulated patients administered in an OSCE(98). Results from knowledge-based assessments have also been shown to increase in a predictable manner with residents performing better than final year students, who in turn perform better than students in earlier years(61). This strong direct and indirect evidence has led to the conclusion that it is critical that student assessment systems include evaluations of students' knowledge(53,57,86).

This leads to the second aspect mentioned above which is whether multiple-choice question really are the most appropriate format to assess students' knowledge. Again, from the perspective of indirect validity, multiple investigations have documented that student scores on multiple-choice question format tests correlate well with scores on other formats designed to measure knowledge, such as open-ended or free response questions(57,99-103). For direct validity, the large number of questions that can be included in this format allows for adequate use of an exam blueprint to ensure content validity. For construct validity, multiple-choice questions are recommended relative to true/false questions as the latter tend to be limited to assessing knowledge that is either categorically right or wrong, or to distinguishing fact from opinion(35,104). From a generalizability perspective, the large number of questions that can be tested in a short time frame ensures that students are offered sufficient opportunities to demonstrate their knowledge and understanding. This differs from formats that are very time consuming such as oral examinations(105-108) or simulated patient encounters(85), where a student's grade may be based on one or two cases. In these latter situations, students' specific experiences, areas of interest, and/or luck influence their performance to the extent that the score obtained on the one or two scenarios tested is not representative of their true knowledge base(50) (see also a later discussion on content specificity). This is one of the main reasons why oral examinations have been eliminated from many high stakes testing programs, such as that of the American Board of Internal Medicine. Finally, multiple-choice questions also have greater generalizability because they can be computer scored and, for example, inter-rater agreement is not a concern.

One problem that has been associated with multiple-choice questions is the impact of cueing on students' performance(99,102,109,110). This cueing refers to the situation where students are prompted towards correct answers by seeing them in a list of short alternatives rather than being required to recall or generate correct answers from their knowledge(102). Although open-ended questions have been proposed as a solution to the cueing effect, such formats are plagued by psychometric problems such as low validity and generalizability(35,110,111). These problems result from the difficulty in writing open-ended questions that clearly indicate the focus and purpose of the question, the detail required in the response, and the level of specificity required(111). Hand-scoring also decreases reliability (because of subjectivity) and feasibility(because of the extended time required to hand-grade the questions). Finally, the generalizability of such open-ended questions is also lower because students take longer to respond to such questions. Therefore, fewer questions can be assessed within a given examination time period.

Three solutions to this cueing problem have been suggested, all of which vary the number of options in the response set. In the first, the number of options varies according to the number of viable alternatives and can range from two (*e.g.,* yes/no) to more than 25(44). Although this response format has been evaluated in detail for multiple-choice questions used to assess more complex problem-solving skills (see later discussion), the rationale for the format is applicable to knowledge-based multiple-choice questions also. This rationale is that, if only reasonable alternatives are included in the option set, students receive less cueing towards correct answers and away from incorrect answers. For example, consider the following knowledge based question:

> Of the following options, which drug is most appropriately used to minimize the hepatotoxicity caused by a serious overdose with acetaminophen?
>
> A. furosemide
> B. naloxone
> C. N-acetylcysteine
> D. penicillin
> E. ranitidine

In this situation, upper year pharmacy students should be immediately cued away from answers D and E since they are easily recognized as drugs used for completely different, specific indications. Option A may require some consideration in case acetaminophen overdoses could cause, for example, pulmonary edema. However, since the stem of the question asks

for a specific treatment for hepatotoxicity, then students should be cued away from option A. The choice, therefore, is really among two viable options rather than five. Listing only the two viable options(or better yet, listing three options including naloxone, N-acetylcysteine, and sodium bicarbonate) would lead to less cueing towards the correct answer since all options can be used in the treatment of overdoses with various medications. A second example demonstrates when the number of response options should be longer than five as all options are viable:

> Nitroglycerin has several effects on the cardiovascular system. One effect is the primary effect, while the others occur as a consequence of the primary effect. The primary pharmacological effect of nitroglycerin is a decrease in which of the following?
>
> A. afterload
> B. blood pressure
> C. cardiac blood flow
> D. cardiac output
> E. heart rate
> F. preload
> G. renal blood flow
> H. stroke volume

It is recognized, however, that in this format, cueing is still possible as students may recognize N-acetylcysteine (in the first example) or preload (in the second example) as being correct, or guess the correct option rather than recalling the response from their knowledge base. However, preparation of multiple-choice questions with variable numbers of responses is a feasible option that theoretically offers improvements to the traditional multiple-choice question format while requiring no extra resources. Unpublished data regarding the development of such question formats to assess the prescribing knowledge of physician assistants in the Canadian military indicates that approximately 6.5 questions can be developed per hour by **untrained** item writers.[4]

Published data from this same study documents the validity of such formats in that physician assistants with higher training scored significantly better on the variable option multiple-choice question than physician assistants with a lower level of training(112). Acceptable generalizability was also obtained when 103 such questions were completed in conjunction with 73 key-features type questions (see later discussion) in a three-hour exam period (Cronbach's α=0.87 for the entire exam, α=0.84 for the 103 variable number of response multiple-choice questions)(112).

The second solution to this cueing effect has also been studied more in relation to assessing knowledge application rather than simple knowledge base and understanding. This is the extended-matching format (*i.e.*, type R and Pick N formats) currently used by National Board of Medical Examiners (NBME) on Steps 2 and 3 of the United States Medical Licensing Exam (USMLE)(104,111,113). In this format, a series of vignettes are written that relate to a similar topic (*e.g.*, antifungal agents)(111). A list of options is then created that is used for all of these vignettes with the number of options commonly varying between six and 25. In the situation of antifungal agents, the list of options would include a list of names of various different antifungal agents. Students are then required to read the vignette and select the single most appropriate response (type R) or multiple responses (Pick N format) for each vignette from the list of options. For the example of antifungals, students could be requested to identify the appropriate agent in a scenario such as: A fungicidal agent that binds to membrane ergosterol, thus altering cell permeability(111). Any single option can be selected once, more than once or not at all. Once the series of vignettes are completed, the student moves on to another series that focuses on a different topic (*e.g.*, therapeutic options) (see references 111 and 113 for multiple examples of this format).

Limited literature has documented that, relative to true/false and standard, five-option response formats, extended-matching formats have a greater reliability and require less testing time(114). NBME also has identified that the use of the specific extended-matching format is very feasible as it is relatively easy to focus on knowledge application rather than recall of isolated facts and that the structure of the option sets allows relatively rapid writing of questions. When working with **new** item writers, NBME has been able to generate 10 usable items per author per hour. Evidence for the former statement about the ease of focusing on knowledge application rather than factual recall is not provided. However, the examples provided in the USMLE literature for the 2001 Step 2 exam clearly demonstrate that such formats can focus on knowledge application(113).

A third solution to the cueing effect is a computer-graded format where students are presented with a long-menu of options (*e.g.*, > 500 options), all of which are numerically coded(115). Students search for the appropriate answer among those on the list and enter the associated code onto the computer-read answer sheet. Although limited literature is available that assesses the psychometrics of such a response format, it appears that this format is valid and reliable. For example, scores on such a format increase with increased training in a family practice residency(115). In this same study, the generalizability of performance on 32 uncued questions was higher than the generalizability on 32 matched multiple-choice questions with the standard five options (α=0.74 versus 0.60). Response time was also considered with the authors estimating that a traditional multiple-choice question would take 50 seconds to complete while the uncued format would take 75 seconds to complete. This short time requirement and the evidence that all students completed 40 uncued questions within the 50 minutes allotted to the test indicate that no substantial increase in testing time is required to obtain reliable results using such an uncued format. Faculty also indicated that the preparation of uncued questions was very feasible as they focused on writing good quality questions rather than developing reasonably wrong alternatives to place as options for traditional multiple-choice question.

Also related to response formats and generalizability is that the single-best answer-type multiple-choice questions are preferred relative to selection-type questions that require complex instructions and answer selection schemes (*e.g.,* K-type questions where the responses are A: answers 1, 2 and 3 are correct; B: answers 1 and 3 are correct; C: answers 2 and 4 are correct, D: answer 4 is correct, and E: all answers are correct). Such complex question-types have been shown to be less efficient, less discriminating and less reliable than the more simple single-best answer-type questions(104).

---

[4]Data calculated from personal knowledge of item preparation times for MHPE thesis. Assessment of CF PA's Knowledge of Authorized Pharmaceuticals..

In summary, given the literature available regarding validity and global reliability, multiple-choice questions do appear to have significant advantages relative to other formats when used for the assessment of knowledge and understanding. Several response options are available to minimize the potential cueing effect of the multiple-choice question format and the specific response option selected depends, at least partially, on the resources available to develop multiple-choice questions.

Although multiple-choice questions with variable or extended options meet the validity and reliability criteria necessary of a format to assess knowledge and understanding of students, the exclusive use of such questions with tasks written to assess only knowledge and understanding can have an undesirable effect on student learning(58,86). If the format and tasks included in an assessment system focus only on knowledge and understanding, students will also focus only on acquiring and understanding facts and concepts(52-55,57). This is particularly true if the tasks required of the students in the multiple-choice questions focus on detailed factual knowledge as opposed to functional knowledge or that which is required by graduates in any number of the employment, educational, or life situations they may experience following graduation. Given that the educational outcomes desired of health professional programs focus on higher levels of cognitive ability, additional formats and tasks must be included in the assessment system to make it congruent with desired educational outcomes.

From the perspective of feasibility, again multiple-choice questions offer obvious advantages relative to other formats for the assessment of knowledge such as open-ended questions, essays, oral exams, or demonstration-type assessments, multiple-choice questions provide the opportunity for the development and administration of assessments on a large scale and via computer-based systems(44,73), both of which are more efficient than alternative assessment formats. It must be recognized, however, that key to all of this literature on the psychometrics of multiple-choice questions is the fact that such questions must be well written. This requirement spans the need to use an exam blueprint to ensure the appropriate sampling of questions (and to prevent a focus on the teacher's favorite topic, or the most recent topic, etc.); the careful phrasing of the scenario, stem, and available options; and external review of questions. Although clear recommendations are available to guide the proper writing of multiple-choice questions(13,35,75,104), actual preparation of extensive numbers of well-written multiple-choice questions is a challenging task.

Finally, the acceptability of using multiple-choice questions from both the students' and faculty's perspective must be considered. Although little work has been published in this area, a recent study from the University of California, Los Angeles, School of Medicine(63) compared, among other factors, students' perceptions of the appropriateness/acceptability of computer-based simulations, standardized patient examinations, attending physician reports, resident reports, multiple-choice question exams, and oral exams. The fourth-year students in the study indicated that they believed multiple-choice question exams were the best format to use to assess their knowledge base. The acceptability of multiple-choice question formats for assessment of knowledge by faculty is a more difficult area to summarize as the perception of faculty may vary according to their educational philosophy and priority placed on teaching. Much of the acceptability to faculty also relates to feasibility and the time required of faculty to prepare and grade examinations. This leads to a summary statement that, in general, if the assessment format proposed is reasonable as to the time required for faculty involvement and meets minimal psychometric requirements, it will be acceptable to faculty.

Based on the available literature, therefore, it appears that the development of an adequate knowledge base is a critical requirement to attainment of competency and that multiple-choice question are the most psychometrically appropriate format to use to assess this knowledge base. The cueing effect of multiple-choice questions can be decreased by either including only reasonable alternatives as options, using extended-matching items, or developing computer-graded, long-option menus. The undesirable steering effect on student learning, however, must be remembered when using multiple-choice question formats to assess knowledge and understanding. As Blake, *et al*., summarized, the challenge should be recognized and multiple-choice question formats administered in such a way as to "maximize the potential benefits, in terms of providing students accurate and comprehensive assessment of knowledge mastery, while avoiding the potential steering effect of the examination"(86). Blake goes onto to suggest the progress testing administration method as one solution for this challenge(86). A second solution is to not rely solely on multiple-choice question-based testing of the students' knowledge base in the assessment system.

When considering the development of an assessment system for pharmacy students, it must be remembered that the cited research and literature comes primarily from the field of medicine. Although consistent with research in the development of expertise in a number of different fields(95), similar research has **not** been conducted on the development of expertise with pharmacists or pharmacy students. What is clear from the available literature is that the formation of an extensive, integrated, structured knowledge base is central to the development of expertise. This dependency on knowledge should be true, regardless of whether the final desired outcome is the provision of medical care, the provision of pharmaceutical care, managing of a practice, communicating with various audiences, or making rational, ethical decisions(14). All of these educational outcomes require students to have knowledge and understanding of associated, fundamental facts and concepts. With the above limitations kept in mind, this knowledge and understanding should be assessed as part of a student assessment system and the most appropriate format to use is well-written multiple-choice questions.

Summary: *Substantial literature exists to support both the importance of assessing a student's knowledge base and understanding, and the use of multiple-choice questions as the most psychometrically appropriate format for assessment of a student's factual knowledge and understanding of concepts. The focus of such formats should be on the assessment of functional knowledge required of graduates and consideration should be given to using a response option of the variable number, extended-matching, or long-menu selection-type to minimize cueing effects. To avoid undesirable steering effects on student learning, such formats should be administered in a "progress testing" manner with repeated administrations of equivalent assessments given over the duration of the curriculum.*

Knowledge of Skills

The second potential use for written assessments using

multiple-choice questions is in the assessment of students' knowledge of how to perform psychomotor or technical skills (*i.e*., from Miller's *knows how* level)(64). Although it is normally presumed that simulations or models are more valid formats to use for such outcomes, several authors have examined the validity of using multiple-choice question format tests for assessment of these skills. Ram, et al, investigated the ability of written tests of both functional medical knowledge and knowledge of medical technical skills to predict medical performance during daily practice(98). Although the students' scores on the knowledge of skills test were significantly lower than their scores on the medical knowledge test, results on both written assessments correlated significantly ($P<0.01$) with results on assessments of actual, daily performance. Similar results were obtained by Van der Vleuten, *et al*., during an evaluation of the relationship between students' scores on a knowledge of technical and clinical skills test[5] to scores on both a general medical knowledge test and a simulated patient-based assessment of technical and clinical skills that was administered via an OSCE method(97).

Their results indicated good reliability (generalizability coefficients of 0.90 or greater) and a strong correlation between students' results on the knowledge of skills tests and their results on the standardized patient-based assessment, particularly for higher level students (correlation coefficient corrected for unreliability for sixth year students = 0.89). Despite these positive results and the greater feasibility of such assessments relative to demonstration-type formats (*e.g*., standardized patients), the authors warn that heavy emphasis on such assessment formats would encourage another different, undesirable approach to the learning of technical and clinical skills. As they suggest, this could result in students focusing "on the mere knowledge instead of practice of actually doing" the skill.

Some of the technical and clinical skills evaluated in the above medical assessments are also required of pharmacy graduates including, for example, history taking, selected aspects of physical assessment and basic first aid/CPR(14). Knowledge of additional skills listed in the AACP CAPE *Educational Outcomes*(14) could also reasonably be assessed via such methods including, for example, preparation of prescriptions, compounding, and production/administration of sterile dosage forms and enteral nutrition products. Alternative methods must be used for the more cognitive skills(such as identification of drug-related problems, management, and provision of drug information) as assessment should focus on the decisions made in these areas rather than the process followed when making these decisions(see discussion below on clinical reasoning and problem solving).

Summary: *A limited amount of literature is available that documents the validity, reliability, and feasibility of using multiple-choice questions to assess students' knowledge of technical and clinical skills. Since the extensive use of such a format could have undesirable effects on student learning, such questions are not recommended as the major format to use when assessing these types of outcomes.*

Application of Knowledge/Clinical Reasoning (Written Simulations)

The above discussion of written assessment formats has focused on the use of multiple-choice questions as the most appropriate format to assess the knowledge base and understanding of students as related to both biomedical knowledge and knowledge of technical and clinical skills. A third use of written assessments must also be considered; assessment of a type of competency from Miller's second level of *knows how*(64). This is the ability of students to apply their biomedical and/or clinical knowledge. For the clinical knowledge, such application is often referred to as problem solving or clinical reasoning. As with the literature on the development of expertise in the health professions, the majority of work in the area of assessment of these problem-solving or clinical-reasoning skills has been conducted in medicine. Early work focused on trials of various assessment formats such as patient management problems(116), modified essay questions(117), portable patient problem packs (P4)(118), and clinical-reasoning tests(119). The most widely used format was the patient management problem, where students were presented with a brief patient description and then the exam proceeded through various steps such as history, physical exam, laboratory, diagnosis, etc. In each step, the students were required to make a decision regarding the actions they would take. Often the questions used latent imaging pens where, when the student selected and highlighted his/her management option, the response of the patient was revealed. Students were not allowed to return to previous decisions regardless of the impact of those decisions on the patient's status. Scoring of these patient management problems considered both the quality of the decisions and the pathway through the clinical problem as indicators of clinical reasoning.

Development of these formats was based on the theory that application of knowledge or problem-solving ability was a generic skill that once mastered, could be applied to any given situation(120). For example, patient management problems were based on this theory by the fact that the student's pathway through the patient's case was weighted heavily in the grade as if there was an ideal, standardized approach to solving all patient's problems. Subsequent research revealed a number of challenges to this generic skill theory. First, related to the validity of using such assessments as measures of problem-solving skills, the evidence indicated that such skills are highly idiosyncratic, with different experts using different strategies to solve the same patient problem(94,121,122).

The process by which a physician or medical student solves a specific problem is largely based on that person's experiences and the structure of his/her knowledge database, including how his/her individual patient experiences are incorporated into the knowledge structures(94,121). For any given patient problem, physicians or students who have not managed similar patients may use more general problem-solving approaches to resolve the case while other physicians/students who have managed similar patients may use a sophisticated form of pattern recognition(94,121). It is, therefore, difficult to assess the accuracy or thoroughness of a student's problem-solving ability because there is no consensus on the "right" way to solve the problem. Patient management problems were particularly susceptible to this problem as the grading focused on both the quality of the decisions and the number and order of decisions made by the student(96,100).

A second challenge relates to both the validity and reliability of these assessment formats. Research documented that problem-solving ability in medicine is highly domain specific, with the ability of a student to successfully solve a problem or manage one case not being predictive of the student's ability to

[5] A sample question from the true-false knowledge of skills test is "when bandaging the knee the first sleeve is applied above the knee(97).

solve any other problem or case, even within the same domain(39,121,123). This inconsistency in performance is referred to as content specificity and it results in very low inter-case reliability. This, in turn, requires that assessments focusing on the problem-solving skills of students include a large number of problems or cases, often resulting in unacceptably long testing times(124). From the earlier discussion on the relationship between reliability and validity, it is clear that the low global reliability of results when a limited number of cases was included in the assessment would lead to invalid results.

One **written** assessment that has been developed to overcome the challenges described above of domain specificity and idiosyncrasy of problem-solving ability is the key-features approach(73,125). In this method, rather than working through multiple phases of a limited number of long patient cases, short cases, or scenarios are presented followed by questions that concentrate on only the decisions critical to the solution of the problem(73,75,125). The following provides an example of such a format and is taken, with permission, from the Assessment of Medical Assistant's Knowledge of Authorized Pharmaceuticals Exam[6](112). The example also demonstrates the principle of using a variable number of response options. Additional examples can be found in Page, *et al.*(74).

> A 27 year old female comes to see you. She tells you that she has been experiencing shortness of breath for several weeks. She has a history of asthma attacks that are normally easily controlled with salbutamol (Ventolin7) inhalations. These attacks usually occur every two to four weeks. This time though her shortness of breath has continued despite using salbutamol (Ventolin7) two puffs four times a day for the last two weeks. On examination there is no intercostal indrawing, no paradoxical breathing, and she is not using her accessory muscles to breath.

What is the most appropriate therapy for this women at this time?

A. add inhaled corticosteroid
B. add ipratropium
C. increase the dose of inhaled salbutamol
D. replace salbutamol with inhaled corticosteroids
E. switch her inhaled salbutamol to nebulized salbutamol
F. start a short course of oral prednisone

From a direct validity approach, the emphasis on decisions made in particular situations and their rationale rather than the processes used to problem solve minimizes the impact of the idiosyncrasy of problem solving(44,73,104,121). From a generalizability perspective, the short case structure and limited number of questions per case allows a greater number of cases and decisions to be assessed within a shorter time frame, thereby addressing the issue of domain specificity and improving the reliability of the assessments(73). Such written assessment formats have been pilot tested as a measure of clinical reasoning/knowledge application/problem solving as part of the national certification process for physicians in Canada(73,126) and are being used as a summative measure of these same competencies in the student evaluation system at Maastricht Medical School in the Netherlands(44).

Reports from these experiences indicate that such assessment formats are valid measures of clinical problem solving/knowledge application as assessed via both indirect(127) and direct approaches(42,73,126,127). For example, in Schuwirth's study 128 fourth year medical students completed two key-features type exams before and after their general practice clerkship. Eight medical expert tutors also completed the post-test. As evidence of indirect validity, the expert tutors scored significantly higher than the students on the post-test, and students scored significantly higher on the post-test than they did on the pre-test. Content validity of such formats has also been assessed in the large scale studies by the Medical Council of Canada (MCC)(126). Fifty-nine key-features questions that had been developed by a test committee of the MCC were validated by 99 external physicians by asking whether the key decisions tested by the question were critical steps that had to be taken to identify and manage the patient's problem appropriately. Strong evidence for the content validity emerged from this study showing that 94 percent of the key-features originally identified were corroborated by the experts. Acceptable generalizability (Cronbach's α=0.8) of such an assessment format has also been assessed with an estimated 40 questions required to be completed in just over four hours of testing time(74).

The impact on student learning has not been directly assessed, although both the Maastricht and Canadian studies have documented that students and faculty perceive that the key-features format accurately tests knowledge application skills and tests different skills than factual knowledge-based multiple-choice questions(126,127). In the Canadian work, 76 percent of over 3,000 graduating medical students responded that the format tested areas that were very critical or critical to practice and 96 percent said that the format used questions that were at the correct level of difficulty(126). Finally, from the feasibility perspective, the key-features approach offers the same types of advantages as traditional multiple-choice questions relative to extended essays, oral exams, or demonstration-based assessments in that the testing time per scenario tends to be shorter (thereby allowing an increased number of scenarios to be tested resulting in increased global reliability), responses are easier to grade, and resource requirements are lower. It, therefore, appears that key-features testing meets the majority of criteria required to be considered as a recommended format for assessing the ability of students to apply their knowledge to the management of patient problems.

In making the above statement, however, it must be recognized that it is not only the scenario and task requested of the student that has changed in key-features testing, but also the response format required of the student. Changes in the scenario and tasks focus on changing the nature of the competency assessed in the question from that of knowledge and understanding to knowledge application or clinical reasoning. Changes in response format focus on improving the psychometrics of the tests by limiting the cueing effects normally associated with traditional multiple-choice questions. As discussed earlier, cueing is the situation where the student recognizes the correct answer among a list of distracters as opposed to being forced to recall the correct answer(102).

In both the Medical Council of Canada and Maastricht work, a number of different administration and response formats have been studied and used in association with key-features testing. The Medical Council of Canada uses a paper and pencil test and has studied three main types of response formats: short, write-in responses; selection of a number of cor

[6]Developed as part of the MHPE thesis, Winslade(112).

rect responses from a short menu of options (2-45 options); and selection of a number of correct responses from an extended list of options (>1300 options in a single list for all questions)(73). Poor reliability results for the latter type of questions led the Medical Council of Canada to remove this type of response format from subsequent testing. Present investigations are focusing on short-menu multiple-choice question with multiple correct answers (*e.g.,* from the following list of options, select up to seven laboratory tests that you would consider important to order for the above patient) or short write-in response formats with single or multiple correct answers (*e.g*., write in the most probable diagnosis for this patient, or list three medications that could be used to appropriately manage this patient's problem).

Although write-in questions were initially hand-graded, students' answers are now entered into a computer program that searches for matches and grades the responses. Such a grading system was developed for write-in questions addressing diagnosis and management only as a reliable system for assessing responses related to acquisition of clinical data could not be developed(73,74). The Maastricht work uses computer-based testing to present the written case scenarios and has evaluated a number of response formats including multiple-choice questions with single or multiple best answers, one open-ended format(that required manual scoring), and two long-menu answer formats(where students type in their responses and the computer searches for similarities and reports these to the student for confirmation of response; there was one such format for single best option questions and one for questions requesting multiple responses from the student)(44).

As with the earlier discussion on the number of options provided in knowledge-based multiple-choice question formats, the number of options presented in the problem-solving multiple-choice question was dependent on the number of feasible responses(ranging from two to eight). The rationale for this varying number of options is that, for example, complex, critical decisions can be limited to two options or can have an extended number of options. The following is an example of a two option response set as taken, with permission, from the Assessment of Medical Assistant's Knowledge of Authorized Pharmaceuticals Exam[7](112).

> A 34-year-old male is carried in on stretcher. He had been in the mess drinking beer with friends when suddenly he developed severe, excruciating chest pain. He is conscious and talking.
>
> His ECG shows third degree AV block. No family history is available. He tells you that he has not had a recent chest injury or trauma.
>
> You begin O_2 and start an IV line.
>
> Suddenly he worsens, becoming pale, diaphoretic, extremely short of breath and his level of consciousness decreases.
>
> You begin Acute Cardiac Life Support therapy. Your differential diagnoses are acute myocardial infarction, dissecting aortic aneurysm, pulmonary embolism and severe unstable angina. Given this, you wonder whether administration of alteplase(Activase7 r-TPA) is appropriate for this patient now.
>
> Should alteplase be started in this patient?
>
> A. yes
> B. no

Similar situations from within pharmacy can also be imagined: consider the scenario where a woman brings her child, who suddenly developed a very high fever, is lethargic and complaining of headache, to your pharmacy. The question as to whether you recommend that the mother take her child to the physician immediately has only one of two answers, but it is indeed a critical and complex question with important long-term ramifications. From outside the health professions, consider the example of the question: do you take this person to be your lawfully wedded husband (or wife)? Again, a critical, complex decision with incredible long-term ramifications and only two options! At the other extreme, when asking students to select which antihypertensive agent they would recommend a physician change to for a woman experiencing nifedipine-induced ankle edema, the list of options could be four, seven, ten, or fifteen. The number depends on the specific scenario (*e.g.,* what other medications she is on or has tried, or other medical conditions) and should be guided by the number of reasonable alternatives.

Validity results from Schuwirth, *et al*., multiple studies evaluating these different response formats indicate that there is a strong correlation between student performance on multiple-choice questions, open ended, and long-menu answer formats and that performance, in general, on all types of response-format exams increases with increased training(44). Acceptable reliability was possible to attain with all the response formats, although the open-ended and long-menu formats required significantly longer response times per question than multiple-choice questions. This resulted in a need for longer testing time to accommodate the number of questions required to attain adequate reliability. This led to challenges with the feasibility and acceptability of such response formats.

The summary from Schuwirth was that no firm recommendation could be made about the superiority of any single response format and that marginal gains in validity with open-ended or long-menu formats had to be balanced against the greater resource requirements associated with these formats. From the combined results of the Medical Council of Canada and Maastricht work, however, it does become clear that open-ended or write-in questions **that require manual scoring** are unacceptable for use in large scale, high stakes evaluations as they have a lower reliability and substantially greater resource requirements for grading. These summary statements are also supported by work from the NBME(111). The more desirable response formats, therefore, include multiple-choice questions with variable number of options presented or computer-scored (and ideally computer-administered) long-menu selection formats. The advantages and disadvantages between single best answer and multiple answers remains unclear.

To perhaps add even more confusion to the issue of the ideal response format, the third option of extended-matching items as used by the NBME must also be considered(111). This format, which follows specific guidelines for all of the scenario, stem, task, and response options, is presented by NBME as a way to both minimize cueing and assess knowledge application rather than simple knowledge recall. As discussed earlier, a series of scenarios are written that are linked to an extended list of options related to a specific topic. Students are asked

[7]Developed as part of the MHPE thesis, Winslade(112).

to select the single or multiple best answers from the extended list for each scenario. Although the NBME stated in a 1993 publication(111) that they prefer to use single best answers because, relative to multiple correct answers, the response format presents a clearer task to the students, has a clearly correct answer, and is easier to score, the 2001 NBME/USMLE Bulletin lists several questions with multiple best answers as part of the sample questions for Steps 2 and 3 of the exam(128). Among the advantages of extended-matching options, NBME states that the questions are easy to write. Comparisons do indicate that this is an advantage relative to the average two hours of writing and reviewing time required to develop each key-features case by **experienced** item writers(127). A second advantage stated is that it is, again, easy to write questions that focus on knowledge application rather than simple factual recall. The examples of the extended-matching items provided in the USMLE 2001 information on Steps 2 and 3 of the USMLE(113) do emphasize clinical knowledge application. These include questions on such decisions as laboratory investigations, most likely diagnoses, and appropriate therapy for individual patient problems.

The question that remains, however, is the comparative usefulness of the extended-matching format relative to the key-features format for the assessment of clinical problem solving. No literature is available that compares the theoretical base for NBME's extended-matching format with the key-features format, and no comparative studies have been published. From the limited information available, however, it does appear that the key-features format focuses to a much greater extent on the decisions that have been identified as critical to the appropriate management of the specific patient problems selected from the general objectives for inclusion in the exam blueprint(129).

Theoretically, this should improve the validity and reliability of the results as these critical decisions should be agreed upon by experts more easily than, for example, the general process that should be followed when managing a specific patient(see above discussion on idiosyncrasy of problem-solving approaches). This issue has been studied by Bordage, *et al*. in their content validation studies(126). Results indicated that 94 percent of the key features originally identified were corroborated by external reviewers. This type of research, and the transparent, evidence-based approach to the development and testing of the key-features format, with peer-reviewed publication of all key results, lends support to the use of such key-features formats.

In applying the above information on key-features testing, extended-matching items, and the various different response formats to the assessment of pharmacy students, it must again be remembered that this literature comes almost exclusively from the field of medicine. No information is available regarding the use of key-features type testing to assess the ability of pharmacy students to apply their knowledge to manage either patient (*e.g.*, pharmaceutical care or drug information) or non-patient (*e.g.*, management or drug distribution) problems. The similarities between clinical reasoning in managing medical problems and managing drug-related problems could support the application of such literature to the assessment of knowledge application in this area of pharmacy. This type of assessment is also theoretically applicable to a range of decisions that demonstrate an ability to apply knowledge and understanding, including those beyond patient management decisions. Scenarios and decisions related to additional professional practice-based and general ability-based outcomes required of pharmacy graduates such as practice management, the provision of drug information and education, and communication could be developed to assess the level of knowledge application expertise of students in these outcomes. Problems requiring application of more basic biomedical and biopharmaceutical knowledge could also be developed and used to assess students' higher levels of cognitive skills in these areas. Such formats would require evaluation via carefully designed pilot projects in order to determine their validity and reliability as measures of knowledge application expertise and their impact on student learning, feasibility, and acceptability to both faculty and students. This same statement is applicable to the use of extended-matching formats for the assessment of knowledge integration and application.

Summary: *Sufficient literature exists to support the use of key-features questions as one psychometrically appropriate format for assessment of a student's clinical-reasoning skills or ability to manage/solve individual patient problems. The focus of such formats should be on the critical decisions required to appropriately manage a specific patient's care. There is no clearly superior response format to use with such key-features questions. Alternatives include variable numbers of options; computer-graded, write-in responses for diagnosis and treatment-related decisions (ideally in conjunction with computer administration of the key-features question); or extended-matching formats. Some authors support the single best answer option relative to multiple correct options, but little empiric support is provided for advantages of either system. Therefore, the response format selected should be guided by both the nature of the content of the question and considerations regarding resource requirements.*

As with the recommendation relating to the use of multiple-choice questions for assessment of knowledge, the advantages of a progress testing administration of key-features assessments should also be considered. Although no literature is available that examines such a system, Schuwirth recognizes that such a process would theoretically encourage a more continuous learning by students(44). In addition, to maximize the usefulness of data collected via this format for quality assurance, common assessments should be developed for use by multiple colleges and schools.

Finally, it must be recognized that new developments in the assessment of knowledge application and clinical reasoning are occurring at a rapid rate. As researchers advance the understanding of the changes in knowledge and knowledge structure that are associated with increasing expertise, new assessment formats can be developed that match these changes. One such format is the script concordance test that is based on assessing whether the knowledge of examinees is appropriately and efficiency organized to be useful in clinical decision making(130,131). This format uses the key-features approach to identify the critical decisions required to manage specific patient scenarios, but alters the task and response format. Preliminary results indicate that such a format has acceptable validity and reliability, and since the response formats can be computer graded with simple programs, the format requires fewer resources to develop relative to other formats that require complex computer scoring systems. Since this format is based specifically on theories about the development of **medical** expertise, and in particular expertise in medical diagnosis, the applicability at this time for pharmacy students is questionable.

COMPUTER-BASED SIMULATIONS

Although the above discussion and recommendations relate to

the use of **written** formats to assess knowledge base, understanding, application, and problem solving, any analysis must also consider other formats that are available which aim to assess similar outcomes. In the first two fields of knowledge base and understanding, few alternative formats are available the focus specifically on these areas. Although many formats implicitly address knowledge base and understanding, the focus of these formats (such as oral examinations and simulations) has been much more on assessing clinical skills(51). It is for these reasons that multiple-choice questions are so widely recognized as the most appropriate format for focused assessment of a student's knowledge base and understanding. In the areas of knowledge application and problem solving, however, several additional formats must be considered including simulations and demonstration-based assessments.

Specific simulation formats include computer-based simulations, models, environmental simulations, and simulated patients. The first format, which has been developing over approximately the last 30 years, most closely aims to assess the same types of problem-solving skills as the key-features formats discussed above(51). Such computer-based simulations (CBX) are presently being studied by the NBME in the USMLE as a format to assess clinical decision-making skills of medical graduates in a more realistic and integrated way than via written case-based assessments that use the extended-matching response option(128). NBME is careful to emphasize that the CBX format measures application of medical knowledge skills and **does not** measure skills that require human interaction such as history taking, physical examination, education, and counseling, professionalism, and humanism. Although the original computer-based simulations were basically patient management problems that were administered via computer, present day versions present students with sophisticated simulations of patients and care settings and require them to select options from a full range of patient care decisions.

Despite these advances, the majority of actions required of the student are still limited to providing evidence that(s)he *knows how* to perform a task rather than *showing how* they would perform the task. The exceptions to this statement include simulations that, for example, require students to write a sample chart documentation or, possibly, to interpret a multimedia presentation of a physical finding such as an ECG. Other situations such as interpreting of multi-media presentations of heart sounds still do not require students to *show how* they perform the cardiovascular exam - but are limited to requiring them to demonstrate their ability to interpret a specific physical finding.

In the computer-based system developed and being implemented in the Step 3 of the USMLE by the NBME, nine cases are administered via a complex software administration and scoring system(Primum7) and students are allowed four hours to complete the nine simulations[8](113). The cases follow a branched scenario (meaning students must work through the entire case with multiple options offered at multiple times through out the patient scenario). Free text entry of orders is the main method of interacting with the computer-generated standardized patient and environment, and students can advance time as they feel appropriate to perform follow-up activities. With the computer-based simulations, the student is expected to diagnose, treat, and monitor the patient's condition as it changes over time and in response to treatment(63). The computer software records orders, decisions, time changes, etc., and a complicated grading system is used to assess the students' performance. This grade includes a consideration of the process followed by the student when managing the patient and counting correct actions taken at appropriate times, as well as harmful actions(63).

Limited information is available in the peer reviewed literature about the psychometrics of the current computer-based simulation format although it has been tested for several years. Edelstein, *et al.,* compared student performance on several assessment formats including on ten sample NBME computer-based simulation questions, eight simulated patients, and scores on USMLE Steps 1 and 2 (multiple-choice questions that focus on knowledge and knowledge application)(63). As anticipated, correlations (uncorrected for unreliability) among performance on the different assessments were intermediate, ranging from a high of 0.84 between the USMLE Steps 1 and 2, to an intermediate of 0.40 between the computer-based simulations and the USMLE Step 2, to a low of 0.24 (0.4 corrected for unreliability as by authors) between the computer-based simulations and standardized patient evaluations.

The authors recognize the challenges associated with interpreting such intermediate correlations and make the general statement that the results lend support to the theoretical construct that the three formats (multiple-choice questions in USMLE Steps 1 and 2, computer-based simulations, and simulated patients) "are measuring different, albeit related, domains of competency." Reliability was calculated via Cronbach's alpha and was 0.54 for the computer-based simulations and 0.69 for the simulated patient evaluations. It is important to note that these values are substantially lower than the 0.80 normally expected of formats used in high stakes testing and the authors recognize that an increased number of scenarios would be necessary to increase this reliability.

This is an interesting recommendation in view of the fact that the NBME Step 3 exam is currently using only nine cases in the computer-based simulations component. Regarding educational impact, Edelstein also compared student perception of the value of the above formats, plus attending physician assessments, residents' assessments, and oral examinations as measures of various different components of clinical competence. As reported earlier in this paper, the students believed that the USMLE Step 1 and 2 multiple-choice questions were the best format for assessing knowledge base, while the computer-based simulation was the best format for assessing clinical decision-making skills. Residents' reports were perceived to be the most appropriate to assess the student's overall ability as a doctor(computer-based simulations rated third behind attending physician's reports and before oral exams, standardized patient exams, and multiple-choice questions). Feasibility was not addressed in the study.

Since both computer-based simulations and key-features testing aim to assess the clinical-reasoning/problem-solving abilities of students, it would be useful to compare the psychometrics, performance, and impressions of students of the two formats. No such studies have been completed to date, so any comparisons made must be based on theory. From a fidelity(or face validity) perspective, computer-based simulations should offer advantages by more realistically representing the clinical

[8]Step 3 is administered over two days (14 hours of testing time) and is made up of 500 multiple-choice questions that are to be completed in 10 hours of testing and nine simulated patients that are to be completed in four hours of testing time. The 2001 Exam Bulletin states that the candidate's performance on the simulated patients will affect the candidate's score and his/her pass/fail decision - but that the weight assigned to the simulated patient is not greater than the time allotted to the questions relative to the multiple-choice questions.

decision-making situations required of medical graduates.

Regarding direct validity, both formats are based on assessing the ability of students to apply their knowledge to the resolution of a patient's problem through accurate diagnosis, treatment, and care. The key-features approach focuses solely on the decisions made by the students, however, while the computer-based simulations also considers the process followed by the students when making their decisions.

Theoretically, this offers validity challenges similar to those experienced with the patient management problems where it was recognized that different experts used different approaches or processes to solve similar problems. This focus on process also offers reliability challenges related to consistency of scoring of responses by multiple graders. Finally, the emphasis on process and the complete management of the patient situation results in the computer-based simulation cases requiring substantial longer time periods to complete (NBME allows just under 30 minutes for completion of one case). This time requirement obviously limits the number of patients and scenarios that can be assessed during a reasonable time period.

As discussed earlier, reliability is lower when the number of cases tested is small. Low reliability with a low number of cases is a particular problem when assessing students' problem-solving abilities due to content specificity (see earlier discussion in section on key-features testing). Between the two formats, little difference would be expected in the impact on student learning or acceptability, especially if the key-features format is administered via computers. Finally, from the perspective of feasibility, the key-features format should offer substantial advantages as the format can be administered without the need for development of complex software to administer and score the exam. Even computer-administered key-features testing should require fewer resources to develop and implement than sophisticated computer-based simulation formats.

The latter consideration of feasibility is particularly important when considering the use of computer-based simulation formats in the assessment of pharmacy students. Few simulations have been developed and tested as measures of complex pharmacist-specific responsibilities(132). This is reasonable in view of the fact that roles such as the provision of pharmaceutical care are relatively new to the profession and controversies remain within the profession regarding the standardized patient-specific tasks and responsibilities associated with the provision of such care(133-137). This has resulted in the situation where no research has been completed that examines the detailed changes in knowledge and skills that occur as pharmacists develop expertise in such roles.

Without this information, it is a difficult, and perhaps premature, step to dedicate substantial resources to the development of computer-based simulations that aim to assess student performance in these roles. These comments should not be interpreted to mean that computer-based simulations designed for different health professionals are not useful for pharmacy students for teaching and learning purposes. However, it would be inappropriate to include such simulations (that focus on, for example, history, physical exam, diagnosis, and management from the perspective of a **physician)** in a summative assessment system for pharmacy students. Finally, given the ever-increasing pace of development in computer simulations and assisted learning, development of such assessment formats may become more feasible, and their use more reflective of higher levels of performance, in the near future(138).

Summary: *There is limited, preliminary information to support the validity, reliability, and student acceptability of computer-based simulations as a format to assess medical students' knowledge of clinical decision making. The feasibility of using such a format, however, is questionable when the fact is considered that the NBME has invested over 20 years of research and development into this format and has only this year included the computer-based simulations as a scored component of the USMLE Step 3 exam. Similar sophisticated computer-based case simulations that focus standardized patients specifically on assessing pharmacists' abilities to manage patient drug-related problems or apply other pharmacy-specific knowledge to the identification and management of problems are not available. This current lack of feasibility lends support to the use of either written or computer-administered key-features formats, rather than computer-based simulations, for the assessment of knowledge application/clinical reasoning/problem solving.*

MODELS AND MODEL ENVIRONMENTS

Despite the advances made in computer-based simulations, it is clear that such formats do not, at present, provide an opportunity to assess students on their ability to perform skills that require widely integrated tasks or human interaction, including those that require demonstration of general ability-based outcomes such as verbal communication and professionalism. The use of models moves toward evaluation of these types of outcomes that are related to students' ability to *show how* they would perform a task(64). Models available range from relatively simple ones for learning technical or physical examination skills (*e.g.*, anatomical models for learning gynecological examinations)(78) to more complex forms that focus on more integrated skills and provide feedback to students on the accuracy of their performance (*e.g.*, artificial arms for training of venipuncture, models used in CPR training). The main use of such models in medical education has been for teaching purposes as they provide an opportunity to practice part or complete skills(90) and offer opportunities for the provision of formative assessment(78).

In the summative assessment situation, models are less frequently used except for evaluating the basic psychomotor or clinical skills required of students within the early years of the curriculum(58,139). Since their use does not usually reflect the final, integrated skills required of graduates, focusing on such models in final assessments lacks validity and fidelity and may encourage students to limit their practice of the fully integrated skill on simulated or real patients. Consideration could be given to including models in an OSCE-administered assessment as this would decrease the costs associated with the use of more sophisticated simulations such as standardized patients. The potential for decreased validity and undesired educational impact, however, have also limited the use of models in these situations (personal communications, J. Van Dalen, PhD, Director Skills Training Laboratory, Maastricht Medical School, October, 2000).

From the perspective of a system to assess the achievement of pharmacy students, the role of models has received little attention. The use of models has been limited to teaching students how to use such models during patient counseling (*e.g.*, how to use gynecological models to demonstrate proper insertion techniques for diaphragms, or the use of model inhalers for teaching of appropriate inhalation techniques). This use may change as pharmacists become more involved in, for example, blood sampling or blood pressure monitoring as the scope of responsibility expands to therapeutic outcomes

management. In these situations, it would seem reasonable that the training of pharmacists for these clinical skills would follow the principles employed by physicians in their related training. This, in turn, means that the use of models in assessment would be limited to the very preliminary stages of training.

The reason why models have not been traditionally employed in the education and assessment of pharmacy students could be because the psychomotor and clinical skills traditionally associated with pharmacists have focused either more on non-direct patient care activities (such as compounding and dispensing) or clinical skills (such as communications). The former are usually learned and practiced in a simulated environment (*e.g.,* laboratory based courses that, in the final years, occur in simulated pharmacies), while the latter focus more on role playing and, more recently, on the use of simulated patients (see following discussion).

Theoretically, simulated environments offer a good opportunity for assessment of students, especially when the task or responsibility required of the student in the simulated environment matches well with the tasks required in real practice. Such is the case for compounding and dispensing where students can be presented with prescriptions, raw materials, and tools that are identical to those found in real practice. Therefore, the fidelity of the testing environment and associated tasks is high. However, it must still be remembered that assessment in such a situation represents competence assessment rather than performance in real life(see earlier discussion). Other theoretical advantages and disadvantages to assessment in simulated environments are similar to those encountered during assessment of students in practice environments during experiential education rotations and these are discussed in the section of this paper on Observation-Based Ratings.

Despite the above stated theoretical advantages and disadvantages for assessment in simulated environments, no literature could be located that documented the development of psychometrically-acceptable formats or systems for assessment of, for example, compounding or dispensing skills in simulated environments. Even searches for systems from other professions or careers that could theoretically use such systems (*e.g.,* laboratory opticians who grind and prepare lenses according to prescriptions, radiation technologists, or laboratory technicians) did not provide such documentation. Often personnel for these positions appear to have varying educational requirements and the literature that was available focused more on certification programs rather than educational programs(140,141).

Summary: *There is, at present, little literature available documenting the use of models for assessment of the complex skills required of health professionals. In medicine, the use of such models appears to be limited to assessment of very preliminary skills in early years of the curriculum. No literature is available detailing the development of psychometrically acceptable systems that focus on assessing students in simulated environments such a compounding or dispensing laboratories. Despite this lack of information, there are theoretical strengths and advantages of using such systems, including high fidelity and feasibility. Principles from assessment formats used in experiential education rotations could be applied to the development of assessment systems for use in these simulated environments.*

SIMULATED PATIENTS

As discussed earlier, there is often a presumption made that assessment formats that require students to *show how* to do a task are better predictors of future, real life performance than measures *of knowing* or *knowing how*(64). This presumption is particularly strong for standardized patient assessment formats that are administered via an OSCE. Despite the centrality of this presumption to the development of rational student assessment systems, little research has been published that addresses this issue. The most thorough investigation was recently completed by Ram, *et al.*,(67,98,142,143) in his PhD thesis entitled "Comprehensive Assessment of General Practitioners: A Study on Validity, Reliability, and Feasibility." In this research, Ram compared the ability of physicians' scores on written assessments of both general medical knowledge and technical/clinical skills, and their scores on demonstration-based assessments using standardized patients administered via an OSCE, to predict their scores on assessments of real performance in daily practice. Results did not support the general assumption that assessments that required students to demonstrate competence were better predictors of actual medical performance than non-demonstration-based assessments(98). In fact, students' results on carefully constructed written examinations(objective format) correlated with actual medical performance as well as, or better than, results with standardized patients administered in the OSCE(written exams to medical performance Pearson correlation corrected for unreliability 0.43-0.56, standardized patients to medical performance Pearson correlation corrected for unreliability 0.33-0.59).

Although these studies are among the few that directly evaluate the predictive ability of demonstration versus knowledge-based assessments for performance in real life practice, support for these findings has been reported by multiple researchers who have shown that students' results on selection-type written assessments of knowledge correlate strongly with their results on simulation-based examinations(39,96). Given these correlations it would be logical that both types of assessments could predict actual performance to the same degree. As Van der Vleuten and Swanson(77) point out, however, and as discussed in the section on indirect validity, the fact that performance on written assessments of knowledge correlate with performance on demonstration-type assessments does not necessarily mean that the two formats are assessing the same competency or construct. It could be that they are measuring related but different concepts. Given this difficulty in interpreting correlational results, it is generally recommend that standardized patients should be incorporated into systems to assess student achievement(77). This is because, apart from validity, simulations with standardized patients that are administered via an OSCE format offer distinct advantages relative to other forms of assessment.

Two of the most important advantages of standardized patients, regardless of whether they are administered as part of an OSCE or in other administration systems, are their high face validity (or fidelity) and impact on student learning. Students feel that such assessments require them to perform tasks that are closely linked to those that will be required of them in real life(63,78,85,144-146). They, therefore, see the relevance of such assessment formats. More importantly, however, is that such formats are congruent with the educational outcomes desired of graduates: the format requires students to demonstrate their abilities to fulfill the final educational outcomes in an integrated manner. This congruency is necessary if students are to be encouraged to undertake integrated, deep, conceptual learning and to practice their complete, integrated skills(52-55,147). This focus on integration of outcomes also provides

an opportunity to assess the more general abilities of graduates within the context of performance of professional tasks.

Similar to the AACP CAPE *Educational Outcomes* required of pharmacy graduates(14), many health professions have identified a series of general abilities that are both expected of university graduates and required for successful fulfillment of professional practice-based outcomes(27,148-150). These include abilities related to communication, critical thinking, ethics, and social responsibility. Many of these abilities are linked together in outcomes associated with professionalism and/or humanism(151-153). Although some attempts have been made by the health professions, and in particular nursing(29), to assess these more general abilities as independent characteristics, and separate from the performance of professional practice-based outcomes, the majority of literature supports the theory that these general abilities are too contextually bound to be assessed as independent characteristics(19,91,92,154,155). This movement is, therefore, similar to the shift from attempts to assess problem-solving skill as an independent characteristic towards its assessment in the context of decision making for a patient in either a simulated or real-life situation.

Numerous articles address the efforts to assess these general abilities within the context of professional competencies, particularly in the areas of the assessment of communication skills and professionalism/humanism (including ethics and social accountability)(156-159). In a review of the use of standardized patients by two noted assessment experts(77), one of the key conclusions drawn is that standardized patient-based tests should be used to assess integrated clinical skills with a focus on history taking, physical examination, and communication skills. Although not specifically stated in the conclusions, it is clear from the review and the articles upon which it was based that general abilities related to professionalism (*e.g.,* attitudes and ethics) are integrated within the areas of history taking and communication skills(160,161). Support for this recommendation was also given by the Consensus Statement of the Researchers in Clinical Skills Assessment on the Use of Standardized Patients to Evaluate Clinical Skills(162).

This focus on using standardized patients to assess primarily integrated, complex skills remains controversial however, especially when considering the ideal station format in a standardized patient-based OSCE. In some OSCEs, a couplet station format is used where the standardized patient presents in the first half of the couplet station and the students are required to answer either verbal or written questions related to the standardized patient in the second, couplet station(79, 80). Although some reports indicate that such formats provide unique assessment data(163) other reports do not support this finding(164). Van der Vleuten and Swanson(77) strongly recommend that standardized patient-based assessments, and OSCEs in particular, focus on hands-on, clinical skills and ***not*** include couplet stations with written questions to assess knowledge and understanding as there are more appropriate testing formats to use for the latter. Given the limited amount of literature concerning this issue, however, groups continue to use different formats for OSCEs with the Medical Council of Canada using couplet stations(80), while the British Columbia College of Pharmacists(24), Canadian Optometry Examining Board of Canada(32), and NBME(165) using or proposing the use of hands-on stations only.

Although the use of standardized patients to assess students' skills offer several advantages relative to other assessment formats, a number of challenges exist related to their use. The majority of these challenges relate to the feasibility of developing a way to use these standardized patients that provides acceptably reliable results.

The first is the need for a relatively large number of simulated scenarios, and the ensuing needs for long testing times and many standardized patients(77,166,167). These needs relate to the domain or content specificity of medical performance discussed earlier where the ability of a student to successfully manage one case is not predictive of the student's ability to solve any other problem or case, even within the same domain(121,123). Therefore, to gain a representative, reliable sample of the student's ability requires a minimum of 15-20 scenarios/stations(77,167,168). This requirement obviously supports the use of an OSCE administration when one-time-only or high stakes testing is being considered. Via this administration, 15 to 20 patient stations can be tested in three to four hours depending on the nature of the skill tested and the duration of each station. Alternatively, if standardized patients are being used over a period of time to assess student performance(81,84,85), then each student should be offered the opportunity to be assessed 15-20 times on their performance with different standardized patients in different content areas. Given the class size for most health professions programs, this latter requirement offers on almost insurmountable challenge. Basing students' grades, or a significant portion of their grades, however, on their performance of complex skills with only a few standardized patients is not defensible from a reliability perspective.

A second challenge related to reliability is the need for trained raters to assess students in each of their performances with standardized patients. Given that such demonstration-based assessments require a judgment of quality of performance, two questions arise related to raters. First, who should rate the students' performance and second, what tools should be used to rate the performance. For the first question, a number of investigations have evaluated the validity and reliability of using the standardized patients themselves as raters relative to using faculty/expert practitioners(167,169,170). In general it appears that both groups can provide valid, reliable assessments. As Van der Vleuten and Swanson(77) point out, practical and educational issues may be the most important factors to consider when determining whether standardized patient or faculty should be used as raters.

Certainly it is more feasible to consider using the former and standardized patient may be able to offer an appropriate perspective on the more professional/humanistic outcomes (such as communication skills or ethics) or on the acceptability of certain technical or physical examination procedures. On the other hand, faculty should be more capable of assessing outcomes such as clinical reasoning, diagnostic and therapeutic decision making, and technical accuracy of history and physical examination techniques. The best situation might be one where faculty provide summative and formative assessment from a professional perspective while standardized patients provide assessment regarding the more humanistic/professional outcomes. In situations of limited resources and/or high volume, however, it would be appropriate to use well-trained standardized patients to provide both formative and summative assessments(77,79,165). What is clear from the literature is that no more than one rater is required for summative assessment: if more raters are available they should be used to increase the number of stations assessed rather than having two raters completing summative assessment in one station(77).

The second question about raters involves whether

detailed checklists or global rating scales should be used to assess students. The former are station specific and require raters to consider whether or not students completed each of the specific components on the checklist. As Regehr states, the use of such checklists turns raters into recorders of behaviors rather than interpreters of behaviors, thereby removing the subjectivity and need for professional judgment from the assessment process(171). Global rating scales tend to be station independent and deal with broad categories of outcomes such as history taking, physical examination, overall medical performance, or communication skills(172).

Such global scales require assessors to judge the overall adequacy of performance in these areas on either general scales (e.g., poor, adequate, good, excellent) or criterion-based scales (e.g., for physical examination used the correct techniques in the appropriate sequence; interpreted physical signs and symptoms appropriately). Although the initial research favored the use of detailed checklists, this effort was based on the assumption that increasing the objectivity of an assessment format would always result in a more reliable and valid assessment format. However, Van der Vleuten, Norman, and De Graaf's reviews of the literature addressing this assumption clearly demonstrate that increased objectivity does not necessarily result in increased reliability or validity(45,46). This is primarily because such detailed checklist formats reward thoroughness as opposed to competence, and do not recognize that, depending on the students' knowledge and experience, different approaches may be used to arrive at correct responses(94,122). This has resulted in a return to the use of more global, less detailed, methods for assessment that require assessors to judge the quality of the student's work or performance rather than simply to record the presence or absence of specific details(48,49,172-174).

This use of such global scales is dependent on the assumption that the skills being assessed are complex and integrated in nature as opposed to specific, step-by-step psychomotor or technical skills. One factor must be considered, however, when contemplating the exclusive use of global rating scales. This is that, depending on how the global assessments are written, they may not provide detailed formative feedback that students can use to improve their performance. For example, rating a student's biomedical knowledge on a four point scale from the low of **limited and segmented** to a high of **comprehensive and well-integrated** does not provide information that is particularly useful to a student. In most assessment situations other than those that focus completely on summative assessment (**e.g.,** one-time-only high stakes examinations such as for certification or in summaries of a series of assessments), more detail can be provided via either written comments or the use of anchors that more completely describe the levels of performance.

For example, the American Board of Internal Medicine's rating forms include a rating of medical knowledge that ranges from a low of **"Limited, fragmented, poorly organized, and applied knowledge of disease, pathophysiology, diagnosis, and therapy is limited. Insufficiently motivated to acquire knowledge"** to a high of **"Extensive and well applied knowledge of disease, pathophysiology, diagnosis, and therapy. Consistently up-to-date. Self-motivated to acquire knowledge"**(175). There is also room to specify what particular areas need attention by the resident and raters are instructed to be as specific as possible in their comments and to avoid global adjectives and remarks such as "good resident" as these do not provide meaningful feedback. The specific type of rating form to be used, therefore, depends at least partially on whether the purpose of the assessment includes formative assessment.

A third challenge associated with the use of standardized patients in primarily the OSCE administration is the ideal duration of time allowed for students to complete their interaction with the standardized patient. In general, the time required is dependent on the nature and complexity of the skill being assessed. However, it is commonly agreed that in an OSCE setting, stations should range from five to 30 minutes in duration(77,176-178). Stations of less than five minutes test trivial issues and more than 30 minutes result in too few stations and content specificity, reliability, and validity become problems(77).

Certainly the most limiting of all of the factors related to the use of standardized patients are the related costs and resource requirements(179). Even schools and colleges of medicine find it difficult to develop their own standardized patient-based assessments that are valid and reliable and, therefore, have built regional alliances in order to make the process more feasible(180).

As with all other assessment formats, the applicability of standardized patient-based assessments to pharmacy students must be considered. Limited information is available in this area, almost all of which addresses standardized patients administered via an OSCE(85,181,182). The only psychometrically rigorous data comes from British Columbia, Canada(24) and the Pharmacy Examining Board of Canada (PEBC) (personal communications, internal documents, June, 2000). The latter group has been pilot testing a standardized patient-based OSCE as a requirement for national licensure and plans to incorporate such a format into its examinations in 2001. PEBC has completed extensive analysis of the use of both non-direct and direct-patient care stations in the OSCE (as consistent with the outcomes required of Canadian pharmacists as defined by the National Association of Regulatory Authorities of Canada)(183), including the development and testing of rating forms. At present, however, no information is publicly available regarding these analyses. The work of Fielding, **et al.** indicates not only that such formats are applicable to the assessment of pharmacists but that results are psychometrically acceptable(24). Results also indicate, however, that although feasible on a large scale, such assessment formats are costly and labor intensive.

On a smaller scale, pharmacy colleges and schools that have developed educational and assessment systems that use standardized patients frequently recognize that feasibility is a challenge that is difficult to overcome(85,181,182). Often a factor critical to the feasibility is an association with an established standardized patient program, usually at a school or college of medicine. Even with these liaisons, it is quite clearly a challenge for individual pharmacy schools to develop standardized patient-based assessment programs that are psychometrically acceptable. This is true regardless of whether the standardized patients are used in an OSCE or as supposedly real patients presenting to pharmacists. This challenge is less of a concern if the standardized patient-based component of the assessment system has a low weight among a number of other assessment formats, but this scenario creates its own problems. If the standardized patients are to be of low weight then it is difficult to rationalize the expense and effort to develop high quality systems. This problem, however, does not arise to the same extent when standardized patients are used as teaching, as

opposed to assessment, tools. In these situations, an exposure to a limited number of standardized patients can be incorporated into a course or curriculum with less concern about reliability.

Summary: *There is a wealth of literature available detailing the development and implementation of small and large scale use of standardized patients for assessment purposes. The majority of this literature deals with standardized patients administered via an OSCE. Substantial evidence supports the validity, reliability, desired impact on student learning, and acceptability of the use of standardized patients. Feasibility, however, remains a challenge and to facilitate the development of such assessments as part of educational programs, regional OSCE administrations are suggested in the literature. Although some controversies remain, such as the specific types of competencies that should be focused upon when using standardized patients, generally accepted recommendations from the literature include that the use of standardized patients should be reserved for assessing integrated outcomes that can not be appropriately assessed via alternative, less expensive formats; a minimum of 15-20 stations should be included that require interaction with standardized patients; stations should be relatively short (5-30 minutes); and summative assessment should be based on global rating scales.*

Less literature is available regarding the use of standardized patients as an assessment format in systems other than the OSCE. Feasibility is limited primarily because of the numbers of exposures to standardized patients that are required for reliable assessment.

OBSERVATION-BASED RATINGS

In the area of assessing students' abilities to actually perform competencies or outcomes in real practice on a daily basis, opportunities are limited to assessing students during experiential education rotations. Beck, *et al*., reviewed the literature relevant to assessment during experiential rotations and recommended that such an assessment system for pharmacy students should focus on the use of observation-based ratings of students' performance, simulations, and written examinations using the extended-matching format(70).

Of these formats, however, only the observation-based ratings assess true performance of students rather than abilities to *know ho*w(written examinations and computer-based simulations) and *show how* (standardized patients administered via an OSCE). Although such ratings are the most common format being used to assess students' performance, Beck, *et al.* and others have identified a number of deficiencies of such an assessment format(70,184-186). These include the fact that most end-of-rotation ratings are completed by preceptors who have actually observed only one or two clinical performances of the student; ratings are strongly influenced by the most recent performance of students rather than a sampling of performance over the rotation; ratings may not really assess the outcomes listed on the rating forms (*e.g*., they are too influenced by items such as personality or communication skills); ratings are insufficiently accurate to discriminate among levels of competency of students; and raters vary in their degree of rating leniency.

A multitude of publications are available recommending methods to improve upon the use and psychometric characteristics of student rating forms(172,185,187-189). From the perspective of using resulting data in quality assurance, however, the American Board of Internal Medicine has completed the most thorough, psychometrically rigorous research(175,190-192). Norcini and Day summarize this research in three categories: content, length, and scale of the rating forms(193). Regarding content they state that "a rating form should include as few items as possible" since raters are infrequently able to distinguish among the competencies underlying detailed questions on a rating form. Similar to the work from assessment of performance with standardized patients, they, therefore, recommend that short, global forms are more useful. In addition, shorter more concise forms have the advantage that it is more likely that they will be completed on a regular basis. Finally, Norcini and Day stated that the decision regarding the scale used should not be belaboured as the impact of selecting three, five, or nine point scales, and anchored versus unanchored scales, is minimal(193).

The same is true for weighting of the areas within the form as none of these factors has been shown to significantly affect the validity or reliability of the use of such forms. Based on these recommendations, the American Board of Internal Medicine has developed and validated two rating forms that are recommended to be used by all residency programs. These are the yearly (summative) evaluation form and the longitudinal evaluation form for use during experiential rotations. These forms require global assessments of the residents' performance in the following:

1. moral and ethical behavior in the clinical setting
2. essential components of clinical competence
   - clinical judgment
   - medical knowledge
   - clinical skills(history taking, physical examination, and procedural skills)
   - humanistic qualities
   - professionalism
   - medical care
3. overall clinical competence as a specialist in Internal Medicine

Some detail is provided on the forms to guide assessments with additional information provided in the Guide to Evaluation of Residents in Internal Medicine(175). Huber, *et al.* have completed the most thorough investigation of the psychometrics of using these forms to assess residents in Internal Medicine(190). From the perspective of indirect validity, supportive evidence includes that rating scores were higher for residents from university-based residency programs relative to those from community hospital-based programs or residents from non-internal medicine programs. This finding is consistent with expectations. Scores were also higher for students who scored better on the associated written American Board of Internal Medicine-in-training examination. Mean inter-rater reliability was good (0.87) but results indicated two problems.

The first problem encountered was that, although the scores were useful for distinguishing levels of performance at the high end, few ratings were received in the unsatisfactory level. Since the residents upon which the study was performed were all expected to perform well, it was not possible to determine if this was an accurate assessment or whether it represented an inability of the rating form to identify unsatisfactory performance. Second, raters could not differentiate among performance in the nine criteria assessed: clinical judgment, medical knowledge, history taking, physical examination, procedural skills, humanistic qualities, professionalism, medical care, and overall clinical competence.

Scores on all categories were strongly correlated and fac

tor analysis indicated that one factor accounted for 76 percent of the variance in scores, with two additional factors accounting for smaller percentages of variance. The authors stated that the factor analysis results support the belief that evaluators may be ranking a global impression of the resident (*i.e.,* a halo effect) rather than assessing each individual criteria. They suggest that, similar to other work in medicine, the three factors may represent biomedical knowledge (cognitive), interpersonal qualities, and technical abilities. The data did not support a finding that bias based on gender or other undesirable characteristics (*e.g.,* personality or undue emphasis on communications) were responsible for the halo effect. To resolve these problems, the authors recommend that either more rater training, or a collapsing of the nine categories into three may improve the validity of the scores.

Although the above study deals with certain aspects of the validity and reliability of observation-based ratings, the use of such formats is still plagued by the problem of insufficient actual observation of students' performances and reliance on the most recent performances of students as the basis for rating (*i.e.,* recency). Hatala and Norman suggest one method of overcoming this problem via the use of encounter cards(189). These encounter cards are pocket-sized forms that require preceptors to rate a student's performance on very broad categories of desired outcomes (see Appendix). Preceptors are required to indicate whether or not they directly observed the student's performance, rate the performance(for summative purposes), and provide specific, behaviorally based comments on the back of the card (for formative purposes). In the trial, students carried the cards and were to request that their preceptors (attending physicians or senior residents) complete a minimum of 15 assessments during the first six weeks of the eight week long general medicine rotation.

Results of this study indicated that such a format was both feasible and reliable, with a reliability co-efficient of 0.79 being obtained for a mean of 7.9 ratings per student (range not provided). Attending physicians were less lenient in the overall scoring and used a wider range of possible scores, with residents' scores ranging only from the three to five category (all satisfactory or above). The authors suggest that training of the raters and a larger sampling of both observed performances and raters could help to address this problem of leniency. However, based on these initial positive results, such an encounter card system has been implemented as a mandatory component of the summative and formative assessment systems for the internal medicine clerkship at McMaster University.

The American Board of Internal Medicine is taking a different approach in attempting to increase the number of resident performances actually observed and rated by preceptors. A current project on the use of mini-clinical evaluation exercises aims to "improve the use of such assessment formats to the level of seamless evaluative activity that requires no complex, structured scheduling"(194). Ideally, preceptors will take 15-20 minutes to observe and rate a resident's history and physical examination performance with four to twelve patients per year in a variety of settings. The short global rating forms use slightly different categories than the observation-based global rating forms(five of the seven are the same as those on the latter form) and similar scales. Based on these assessments, preceptors are encouraged to submit praise cards and early concern notes(175) to identify strengths and weaknesses of residents and to provide formative assessment. Results of the investigation of the use of this format should be available shortly including peer-reviewed publications and, possibly, a policy requiring the mini-clinical evaluation exercises as a component for certification by the American Board of Internal Medicine(194).

In reviewing these forms and the literature upon which they were based, it becomes clear that one of the major advantages of such assessment formats is their usefulness as a measure of a student's performance on multiple, integrated outcomes, including general ability-based outcomes(148). This advantage provides an important link between performance-based ratings and assessments using standardized patients in an OSCE as both are used primarily to assess complex, integrated outcomes. In the ideal setting that maximizes the feasibility and acceptability (to both students and faculty) of an assessment system, these formats should use common global rating forms for summative assessment. This provides students with the opportunity to experience the assessment requirements prior to the high stakes, standardized patient assessment and minimizes the need for faculty to become familiar with multiple rating forms. Finally, the ideal situation for quality assurance purposes would be for a common global rating scale to be developed and used by multiple colleges and schools(148).

In pharmaceutical education, similar work is beginning in the development of such rating scales for both professional practice-based and general ability-based outcomes. For the former, scales developed by both Fielding, *et al.* (24) and the Pharmacy Examining Board of Canada (personal communications, 2000) for assessment of graduates' or pharmacists' performance with standardized patients in OSCEs could be used as a beginning point for development of rating scales useful for assessing all of the relevant professional practice-based and general ability-based outcomes. Hammer, *et al*., recently developed and validated tool to assess behavioral professionalism of pharmacy students also provides an excellent resource for incorporation into a rating scale that addresses all required outcomes(151).

Summary: *Despite a large volume of literature dedicated to the use of observation-based ratings, there is limited literature that documents the validity and reliability of results obtained from any one of the multitude of forms available. General recommendations from the literature to improve the psychometrics include that the rating forms used should be concise and global in nature, raters should receive training, and systems should be developed to ensure a minimum number of documented observations of student performance by preceptors* (e.g., *tools such as encounter cards or rating forms for short observations of clinical skills). A second general recommendation is that there should be less focus on the development of new forms and formats, and more analysis and reuse of available forms and formats.*

Despite the challenges associated with the validity, reliability, and feasibility of effective use of observation-based ratings, the importance of assessing actual performance of students can not be ignored. This leads to the question as to whether there are more psychometrically sound ways to assess the performance of students on rotations. Chart audits/reviews , assessment of videotaped performances, performance with standardized patients who present as real patients during daily practice, and, possibly, portfolios or progress files are such alternate methods. Although chart reviews may offer advantages for certain professions(191), they are less applicable to pharmacy practice as historically, pharmacists have not been

expected to maintain detailed records of patient care or other pharmacy-specific activities. The problem of incomplete documentation, therefore, makes such an assessment format inapplicable for pharmacy students.

Assessment of videotapes of physicians' interactions with patients is currently being used by a number of groups as a format for evaluating real life performance in high stakes situations(143,195). The Membership in the Royal College of General Practitioners (MRCGP) in the United Kingdom includes an assessment of videotaped performances with patients that the member has selected for submission and consideration. This process has been criticized in that it encourages maximal performance to be evaluated rather than average (actual) daily practice(196).

Ram developed an alternative approach where patient visits over one week were videotaped (with consent) and submitted for assessment. Assessors then selected a set number of patient encounters to assess according to a blueprint. This blueprint had been developed based on prevalence of complaints and diseases in general practice and on a nationally accepted description of responsibilities required of a family physician. Results of trials using this process have indicated that the format is valid with acceptable reliability being reached with approximately twelve cases that met the predefined criteria. Acceptance by family physicians was good and the majority felt that their videotaped performances that had been selected for assessment were more representative of their real-life practice capabilities than their performances with standardized patients administered in an OSCE format(142). The costs associated with the videotaped assessment format were also lower than those required of the standardized patients.

Although this format represents a potentially very useful format for the assessment of continuing competency of health professionals, again the applicability to the assessment of pharmacy students on experiential educational rotations is limited. Certainly it would be potentially useful only in sites were the students see patients in a fixed location on a repeated basis (because the videotaping must occur as part of the "regular" activities over an extended period of time, otherwise the student performs differently than during regular practice). This restricts the sites to, perhaps, a consulting room at a community pharmacy or other outpatient clinic. Even in these sites, it is questionable as to whether **all** pharmacy students would see sufficient numbers of patients within a reasonable time frame for the videotaping to provide a valid and reliable sample of actual performance.

Little information is available regarding the use of standardized patients in the assessment of daily performance of health professionals(81). In this situation, a number of standardized patients are scheduled to visit a health professional as supposedly real patients. Rethan's work with physicians in the Netherlands documented that the physicians could not detect the standardized patients even when they knew to expect a visit from a number of such standardized patients within a certain time period(81,82). Limited psychometric data is provided about this format, however, and it would seem logical to assume that a relatively large number of standardized patients would have to be seen by an individual practitioner (12 to 15 from the standardized patient in OSCEs literature) in order to manage the content specificity/reliability problem discussed earlier. Again, this format of using standardized patients as undetectable, real patients for the assessment of pharmacy students during experiential education rotations seems unfeasible. Obviously such standardized patients could not be seen in a hospital setting. If limited to assessment in the ambulatory setting, then such use of standardized patients suffers from the same problems as the videotaped performances in this same setting.

The final assessment format that could potentially be used to assess the actual performance of pharmacy students during experiential education rotations is the log-book, portfolio, or progress files. The use of portfolios to facilitate learning has received substantial attention in the health professions' literature over the years, particularly in nursing education(197-202). More recently, interest has focused on the use of log books and portfolios to assess learning of students in the health professions(198,200-204). In general, portfolios contain two types of information: activities/achievement and evidence for reflection(199). Dennick provides an example of the use of a log book in a medical program that focuses primarily on the former, where students are provided with a list of objectives and specific clinical activities that they should attain/complete by the end of the rotation(202). Some of these must be checked off and signed as completed by the student while others must be witnessed and signed off by a staff member as having taken place satisfactorily.

The log books are used as the basis for a rotation midpoint formative assessment and as a major part of the summative assessment for the rotation. The authors of this study recognize that the primary purpose of such a type of log-book is to "provide structure and focus during the experiential learning cycle...[log books] can consolidate and organize opportunistic learning episodes and they can encourage students to develop responsibility and reflective practice"(202). In this form, however, where the log book emphasizes simple documentation of completion of activities, it is unclear how students are encouraged to develop reflective practice.

Although no psychometric data is provided about the use of these log books for assessment, it is clear that many of the problems inherent with observation-based ratings of students would also apply to the assessment of this type of log-book. In fact, the compilation of the encounter cards, mini-clinical evaluation exercises, or observation-based ratings of medical students or residents that is used to complete summative rotation or year assessments could be considered as a form of such a log book. It, therefore, appears that the use of a log book that focuses on simple documentation and perhaps assessment of performance on individual clinical skills offers little advantage relative to the use of observation-based rating skills.

When discussing the use of portfolios, however, it must be remembered that most portfolios also emphasize reflection. As a very brief explanation of what is meant by reflection (and particularly by reflection on action), reflection requires students to think about a recent experience; relate and compare this experience to their previous experiences, knowledge, and understandings; attempt to develop new knowledge and understandings through this comparison; and then develop plans to test or implement their new knowledge/understandings in subsequent experiential situations(205). This concept of reflection has been linked to two fundamental requirements for attainment of competency as a health professional(201). These are that reflection on action is critical to the development of new knowledge and understanding when learning is occurring via experience as either a student or practitioner(205,206). This ability links to the need for health professionals to be self-directed in their learning and to be able to continue to learn following completion of their formal education (*e.g.,* life-long learners). Second, reflection is also fundamental to deep learn-

ing where practitioners attempt to create new knowledge and understanding by reflecting upon their experiences. This deep learning links to the requirement for health professionals to have a thorough, widely integrated network of knowledge in order to gain expertise (see earlier discussion).

Multiple authors have suggested that one way to facilitate the development of reflection is to dedicate specific time in a curriculum to writing assignments that require students to complete the reflection on action cycle(198-200). Such assignments can also be incorporated into portfolios or progress files where students not only document their activities but include an analysis of their reflection on learning that occurred as a result of the activities. The perceived importance of reflection on the development of professional competence has been recognized by groups such as the English National Board (Nursing) who have requirements for the development of reflective portfolios(207) and the Postgraduate Director of General Practice Education (General Medical Practitioners, UK), who has accepted portfolio-based learning as an alternative means for gaining the required post graduate education allowance for continuing medical education(200).

Although the development and use of reflective portfolios/progress files as a tool to facilitate learning is sufficiently complicated, even more difficult is the use of such formats for the assessment of student achievement(198,199,200,204,208). Challenges occur from primarily the perspectives of validity(199,207), reliability(208), and feasibility(204,208). Burton questions the evidence that supports the validity of using reflection as a means to improve nurses' knowledge and/or outcomes to patients(207). Challis comments that, since a portfolio is a highly individual and unique creation, the assessment must be based on a set of principles that aim to determine if the evidence presented is valid, sufficient, current, and authentic(199). Given this unique nature of the portfolio, reliability of assessment also becomes a problem as indicated by Pitts(208) who used standardized criteria to assess a series of portfolios and concluded that the degree of inter-rater agreement was insufficient for portfolios to be acceptable as a format for summative assessment.

Content specificity could also be a challenge, although the inclusion of multiple examples could help to minimize this problem in generalizability. Finally, feasibility of the use of portfolios for large scale, high stakes assessment of student achievement is also questionable as they, like essay, open-ended, or oral examinations, require "hand" grading that is very time consuming. Finally, several authors have raised the issue that using portfolios as an assessment format runs contrary to the very intended purpose of the portfolio. This is because the portfolio is meant to encourage intrinsic motivation to learn and student responsibility, and requires honesty and the exposure of the student's vulnerabilities and deficiencies. Assigning grades to the portfolio, in turn, substitutes external motivation and perhaps, could encourage students to record experiences and interpretations that they feel will be acceptable to the assessor rather than those which are truly reflective of their experiences and learning. Given these challenges, it does not appear that, at present, portfolios appear to meet the psychometric requirements necessary for an assessment format to be considered as part of an assessment system for the achievement of pharmacy students. However, given the theoretical importance of reflection to the development of professional competence, portfolios continue to be recommended for use as a learning tool and research continues on the development of psychometrically acceptable methods for assessing these portfolios.

Summary: *There is no literature that documents the availability of a format superior to observation-based ratings for the assessment of student performance during experiential education rotations. Portfolios are potentially useful as a means to assess the develop of reflection, an ability that is receiving increasing emphasis as being critical to the development of professional competence.*

RECOMMENDATIONS

Based on the above review, several recommendations can be made regarding the development of a system to assess the achievement of pharmacy students. These recommendations must be placed in context in that the main use for the data generated via such an assessment system would be for the purpose of quality assurance.

1. An assessment system for pharmacy student achievement should include an evaluation of a student's knowledge base and understanding. The most psychometrically acceptable format for this purpose are multiple-choice questions that focus on assessing the functional knowledge required of graduates to fulfill both the AACP professional practice-based and general ability-based educational outcomes(14). The most desirable response format to use with the multiple-choice questions is controversial but consideration should be given to variable number, extended-matching or long-menu selection type to minimize cueing effects.
2. An assessment system for pharmacy student achievement should include an evaluation of a student's knowledge application/clinical-reasoning skills and ability to manage individual patient problems as outlined in the AACP CAPE Educational Outcomes(14). The focus of such a format should be on the critical decisions required to appropriately manage situations relevant to the AACP CAPE Educational Outcomes(14). The most feasible format available that is psychometrically acceptable is the key-features format. Again, however, there is no clearly superior response system to use with the key feature format. Of the alternatives available, the most easily implemented is the variable numbers of options, but consideration should be given to the development of computer-graded extended-matching or long-menu selection type formats. Developments in computer-based simulations for large scale assessment of clinical reasoning should be monitored and advances incorporated into a student assessment system when psychometrically acceptable formats are available.
3. To minimize the undesirable steering effects of the above two written formats of assessment, consideration should be given to administering these formats via centralized, progress testing.
4. An assessment system for pharmacy student achievement should include an evaluation of a student's ability to demonstrate competency in the professional practice-based outcomes required of pharmacy graduates(14). Demonstration of competency in the general ability- based outcomes should be integrated into the assessment of the professional practice-based outcomes(14).

    - Consideration should be given to the use of standardized patients administered via an OSCE for the professional practice-based outcomes that require direct

patient care and direct interaction with other health care professionals. Less resource intensive formats should be used to assess the knowledge and knowledge application associated with these complex professional practice-based outcomes and competency in professional practice-based outcomes not requiring direct interaction.

- Psychometrically acceptable formats for use in simulated environments should be developed to assess a student's competency in professional practice-based outcomes not requiring direct interaction (e.g., compounding or the technical/legal aspects of dispensing).

5. An assessment system for pharmacy student achievement should include an evaluation of student performance in the professional practice-based outcomes during experiential education rotations. As above, demonstration of competency in the general ability- based outcomes should be integrated into the assessment of the professional practice-based outcomes(14). A process should be developed to ensure a minimum number of documented observations of student performance by preceptors (*e.g.,* tools such as encounter cards or rating forms for short observations of clinical skills).
6. Validated global rating forms available from individuals or organizations such as the American Board of Internal Medicine and the Pharmacy Examining Board of Canada should be used as the basis for development of the global rating forms to be used to assess pharmacy students in simulated environments and during standardized patient-based OSCEs and experiential education rotations.
7. Developments in the assessment of portfolios should be monitored and advances that lead to psychometrically acceptable results incorporated into the assessment system. This format could be useful for the assessment of general ability-based outcomes such as self-directed learning within the context of actual performance of professional practice-based competencies.
8. The decision as to what format should be used to assess any specific professional practice-based outcome or sub-outcome, or any general ability-based outcome, should be based on careful analysis of the tasks required of students to fulfill the outcome and the empirical evidence that documents which format provides the most psychometrically sound results for assessment of the required tasks.
9. Given the intended purpose of quality assurance, a goal of the system should be to develop formats and tools that can be used by multiple colleges or schools of pharmacy. Multiple colleges or schools of pharmacy should, therefore, be involved in the development, testing, and refinement of the assessment formats and tools used in the system.
10. Given the number of unanswered questions about the development of pharmacist expertise and the most appropriate formats to use to assess the development of this expertise, the assessment system should contain a research component. Results of studies should be published in peer-reviewed journals in order to provide transparency of the processes used to assess pharmacy students. Research should also evaluate the quality of the assessment system to ensure that the desired educational outcomes and assessment processes used remain appropriate and relevant.

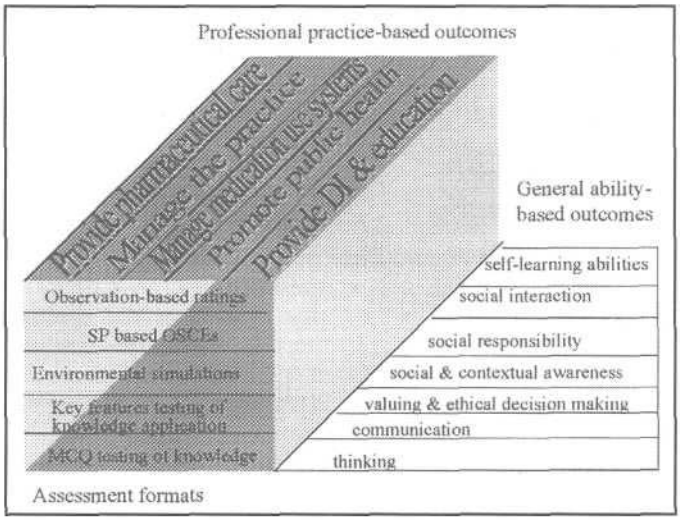

Fig. 6. Example blueprint format on a three dimensional scale(209) (Adopted with permission from the Royal Australian College of General Practitioners).

NEXT STEPS

To develop a blueprint that defines which professional practice-based and general ability-based outcomes should be assessed via a particular format requires a number of steps. First, it is necessary to review the professional practice-based outcomes to ensure consistency of detail. For example, there should be consistency as to whether the need for knowledge and understanding are stated as explicit sub-outcomes or not. The general categories of knowledge required for the outcomes and sub-outcomes should also be listed. For example, whether the outcome requires biomedical, clinical, health care systems, or communications knowledge and understanding. Quite clearly, each of the outcomes and sub-outcomes should have an associated required knowledge base. The contexts in which graduates are expected to be able to fulfill each of the professional practice-based outcomes should also be defined, as well as the expected level of performance. This will ensure clarity and agreement. It is also suggested that the general ability-based outcomes be reviewed and reworded for easier understanding.

Next, the explicit linkages between each of the professional practice-based outcomes and the general ability-based outcomes must be made. This would identify which general ability-based outcomes are required for successful completion of each of the professional practice-based outcomes. This, in turn, defines which of the practice and general ability-based outcomes can be assessed together in an integrated fashion during demonstration-type formats of competency assessment and performance during experiential education rotations. Only after these steps are completed can a blueprint be developed that identifies the weightings of the specific practice and general ability-based outcomes in each of the assessment categories. The type of blueprint that could be developed is best imagined as a three dimensional blueprint as was completed by Fabb for the Royal Australian College of General Practitioners(209). Figure 6 shows such a blueprint for the assessment of achievement of pharmacy students. The x axis represents the professional practice-based outcomes, the y axis the general ability-based outcomes, and the z axis the formats of assessment:

- multiple-choice questions for knowledge and understanding;
- key features for knowledge application/clinical reasoning and problem solving;
- observation-based ratings in environmental simulations (*e.g.,* dispensing laboratories) for demonstration of competency in professional practice outcomes that do not require direct interaction (*e.g.*, compounding);
- standardized patients administered via an OSCE for demonstration of competency in professional practice-based outcomes requiring direct interaction (and integrated with general ability-based outcomes); and
- observation-based ratings in experiential education rotations for performance of professional practice-based outcomes integrated with general ability-based outcomes.

The placing of the general ability-based outcomes on the y axis emphasizes that these outcomes are assessed primarily within the context of performance of the professional practice-based outcomes. Without the completion of these steps, only very general recommendations such as those in this document can be made regarding the best formats to use in a system to asses the achievement of pharmacy students.

Acknowledgements. The author wishes to thank Richard P. Penna, PharmD, Executive Vice President, AACP and Susan M. Meyer, PhD, Senior Vice-President, AACP for their support and interest in the development of this paper and also to Gail D. Newton, PhD, who provided valuable comments and perspective during the writing stages. Finally, expresses appreciation to Lambert Schuwirth, MD, PhD, for his continuing guidance, timely feedback, review, and sense of perspective.

References

1. Boyce, E.G., *A Guide for Doctor of Pharmacy.Program Assessment,* American Association of Colleges of Pharmacy, Alexandria VA (2000).
2. Cave, M., Hanney, S., Henkel, M. and Kogan, M., *The Use of Performance Indicators in Higher Education: The Challenge of the Quality Movement,* Jessica Kingsley Publishers, London. England (1997).
3. Radford, J., Raaheim, K., de Vries, P. and Williams, R., *Quantity and Quality in Higher Education,* Jessica Kingsley Publishers, London, England (1997).
4. Sherr, L.A. and Teeter, D.J., *Total Quality Management in Higher Education. New Directions for Institutional Research Number 71,* Jossey-Bass Inc. Publishers, San Francisco CA (1991).
5. Linn, R.L., "Assessments and accountability," *Educ. Res.,* **29**(2), 4-14(2000).
6. Murray, E., Gruppen, L., Carton, P., Hays, R. and Woolliscroft, J.O., "The accountability of clinical education: Its definition and assessment," *Med. Educ.*, **34,** 871-879(2000).
7. Madaus, G.F., Scriven, M. and Stufflebeam, D.L., *Evaluation Models: Viewpoints on Educational and Human Services Evaluation,* Kluwer-Nijhoff Publishing, Boston MA (1991).
8. Astin, A., "*Assessment for Excellence*," Orynx Press, Phoenix AZ (1991).
9. Hollenbeck, R.G., "Chair report for the Academic Affairs Committee," *Am. J. Pharm. Educ.,* **63,** 7S-13S(1999).
10. *Accreditation Standards and Guidelines for the Professional Program in Pharmacy Leading to the Doctor of Pharmacy Degree,* American Council on Pharmaceutical Education, Chicago IL 1997).
11. World Federation on Medical Education Task Force, "WFME Task Force on defining international standards in basic medical education. Report of the working party," *Med. Educ.,* **34,** 665-675(2000).
12. Boud, D.and Feletti, G., "Student assessment and program evaluation," in. *TheChallenge of Problem Based Learning,* (edits. Boud, D. and Felleti, G.), Kogan Page Publishing, London (1997) pp. 253-259
13. Palomba, CA. and Banta, T.W. *Assessment Essentials: Planning, Implementing and Improving Assessment in Higher Education,* Jossey-Bass, San Francisco CA (1999).
14. Center for Advancement of Pharmaceutical Education, *Educational Outcomes,* American Association of Colleges of Pharmacy, Alexandria VA(1998).
15. *Handbook on Outcomes Assessment,* American Association of Colleges of Pharmacy, Alexandria VA (1995).
16. American Association of Colleges of Pharmacy, Council of Faculties, *Teaching and Outcomes Assessment Committee Report,* Personal Communications, February, 2000 (1998).
17. American Association of Colleges of Pharmacy, Council of Faculties, *Teaching and Outcomes Assessment Committee Report,* Personal Communications, February, 2000 (1999).
18. Gonczi, A., Hager, P. and Oliver, L., *Establishing Competency-Based Standards in the Professions, Research paper No.1,* Australian Government Publishing Service, Canberra, Australia (1990).
19. Hager, P., Gonczi, A. and Athanasou, J., "General issues about assessment of competence," *Assess. Eval. Higher Educ.,,* **19**(1), 3-16(1994).
20. Heywood, L, Gonczi, A. and Hager, P., *A Guide to Development of Competency-Based Standards for Professions, Research paper No. 7,*: Australian Government Publishing Service, Canberra, Australia (1992).
21. Lewis, R.G. and Smith, D.H., *Total Quality in Higher Education,* St. Lucie Press, DelRay Beach FL (1993).
22. Fielding, D.W., Page, G.G., Schulzer, M., Rogers, T. and O'Byrne, C.C., "Assuring continuing competency: Identification and validation of a practice-based assessment blueprint," *Am. J. Pharm. Educ.,* **56,** 21-29(1992).
23. Fielding, D.W., Page, G.G., Rogers, W.T., Schulzer, M., Moody, K.G. and O'Byrne, C.C., "Developing an assessment of pharmacy practice knowledge," *ibid.,* **58,** 361-369(1994).
24. Fielding, D.W., Page, G.G., Rogers, W.T., O'Byrne, C.C., Schulzer, M., Moody, K.G. and Dyer, S., "Application of objective structured clinical examinations in an assessment of pharmacists' continuing competence," *ibid.,* **61,** 117-125(1997).
25. National Commission on Certification of Physician's Assistants, *Content Blueprint for Physician Assistant Initial Certification and 'About Us,'* http://www.nccpa.net/about_us_2.htm (1999).
26. Higgs, J. and Jones, M., *Clinical Reasoning in the Health Professions,* Butterworth Heinemann Press, Oxford, Great Britain (2000).
27. Benner, P., *From Novice to Expert: Excellence and Power in Clinical Nursing Practice,* Addison-Wesley Publishing Company, Menlo Park CA(1984).
28. Paul, R.W. and Heaslip, P., "Critical thinking and intuitive nursing practice," *J. Adv. Nurs.,* **22,** 40-47(1995).
29. Rane-Szostak, D. and Fisher Robertson, J., "Issues in measuring critical thinking: Meeting the challenge," *J. Nurs. Educ.,* **35**(1), 5-11(1996).
30. Fonteyn, M.E. and Ritter, B.J., "Clinical reasoning in nursing," in *Clinical Reasoning in the Health Professions,* (edits. Higgs, J. and Jones, M.), Butterworth Heinemann Press, Oxford, Great Britain (2000).
31. National League for Nursing Accreditation Commission, "*Accreditation standards and criteria for academic quality of postsecondary and higher degree programs in nursing,"* www.accrediting-comm-nlnac.org/2am_stds&crit_fnl.htm(2000).
32. Violato, C, Chou, B.R., McDowell, J.M. and Marini, A., "The Canadian Examiners in Optometry clinical competency examinations: Implementation and psychometric properties," *Can. J. Opt.,* 59(1), 1-5(1997).
33. Boelen, C, Bandaranayake, R., Bouhuijs, P.A.J., Page, G.G. and Rothman, A.I., *Towards the Assessment of Quality in Medical Education,* World Health Organization, Geneva, Switzerland (1992).
34. Newble, D., Dawson, B., Dauphinee, D., Page, G., Macdonald, M., Swanson, D., Mulholland, H., Thomson, A. and Van der Vleuten, C, "Guidelines for assessing clinical competence," *Teach. Learn. Med.,* **6**(3), 213-220(1994).
35. Linn, R.L. and Gronlund, N.E., *Measurement and Assessment in Teaching, 7th Edition,* Prentice-Hall, Inc., Upper Saddle River NJ (1995).
36. Fowell, S.L., Southgate, L.J. and Bligh, J.G., "Evaluating assessment: The missing link?" *Med. Educ.,* **33**, 276-281(1999).
37. Gronlund, N.E., *Assessment of Student Achievement,* Allyn and Bacon, Boston MA (1998).
38. Chalmers, R.K., Grotpeter, J.J., Hollenbeck, R.G., Nickman, N.A., Wincor, M.Z., Loacker, G. and Meyer, S.M., "Ability-based outcome goals for the professional curriculum: A report of the Focus Group on Liberalization of the Professional Curriculum," *Am. J. Pharm. Educ.,* **56,** 304-309(1992).
39. Van der Vleuten, C.P.M., "The assessment of professional competence: developments, research and practical implications," *Adv. Health. Sci. Educ., 1,* 41-67(1996).
40. Cronbach, L.J., "What price simplicity?" *Educ. Meas.,* **2**(2)**,** 11-

12(1983).
41. Ebel, R.L., "The practical validation of tests of ability," *ibid.,* **2**(2), 7-10(1983).
42. Schuwirth, L., Verheggen, M.M., van der Vleuten, C.P.M., Boshuizen, H.P.A. and Dinant, G.J., *Validation of Short Case-Based Testing Using a Cognitive Psychological Methodology [dissertation],* University of Limburg, Maastricht, Netherlands (1998).
43. Cerveny, J.D., Knapp, R., DelSignore, M. and Stier Carson, D., "Experience with objective structured clinical examinations as a participant evaluation instrument in disease management certificate programs," *Am. J. Pharm. Educ.,* **63,** 377-381(1999).
44. Schuwirth, L., *An Approach to the Assessment of Medical Problem Solving: Computerised Case-Based Testing,* Thesis Publications, Maastricht, Netherlands (1998).
45. Norman, G.R., Van der Vleuten, C.P. and De Graaff, E., "Pitfalls in the pursuit of objectivity: Issues of validity, efficiency and acceptability," *Med. Educ.,* **25**(2), 119-126(1991).
46. Van der Vleuten, C.P.M., Norman, G.R. and De Graaff, E., "Pitfalls in the pursuit of objectivity: Issues of reliability," *ibid.,* **25,** 110-118(1991).
47. Van der Vleuten, C.P., van Luyk, S.J. and Swanson, D.B., "Reliability (generalizability) of the Maastricht Skills Test," *Proc. Ann. Conf. Res. Med. Educ.,* **27,** 228-233(1988).
48. Regehr, G., MacRae, H.M., Reznick, R.K. and Szalay, D., "Comparing the psychometric properties of checklists and global rating scales for assessing performance in an OSCE-format examination," *Acad. Med.,* **73,** 993-997(1998).
49. Hodges, B., Regehr, G., McNaughton, N., Tiberius, R. and Hanson, M., "OSCE checklists do not capture increasing levels of expertise," *ibid.,* **74,** 1129-1134(1999).
50. Norman, G.R., "Reliability and construct validity of some cognitive measures of clinical reasoning," *Teach. Learn .Med.,* **1**(4), 194-199(1989).
51. Swanson, D.B., Norman, G.R. and Linn, R.L., "Performance based assessment - Lessons from the health professions," *Educ. Res.,* **24,** 5-11(1995).
52. Frederiksen, N., "The real test bias," *Am. Psych.,* 39, 193-202(1984).
53. Norman, G.R., "What should be assessed?" in *The Challenge of Problem-Based Learning,* (edits. Boud, B. and Felleti, G.), Kogan Page Limited, London (1991) pp. 254-259.
54. Van der Vleuten, C.P.M., Newble, D., Case, S., Holsgrove, G., McCann, B., McRae, C and Saunders, B., "Methods of assessment in certification," in *The Certification and Recertification of Doctors: Issues in the Assessment of Clinical Competence, (edits.* Newble, D., Jolly, B. and Wakeford, R), University Press, Cambridge, UK (1994) pp. 105-125.
55. Van der Vleuten, C.P.M., Dolmans, D.H.J.M. and Scherpbier, A.J.J.A., "The need for evidence in education," *Med. Teach.,* **22**(3), 246-250(2000).
56. Van Berkel, H.J.M., "Assessment in a problem-based medical curriculum," *High. Educ*., **19,** 123-146(1990).
57. Norman, G.R., "Assessment in problem-based learning," in *The Challenge of Problem-Based Learning, 2nd Edition* (edits. Boud, B. and Felleti, G.), Kogan Page Limited, London, UK (1997) pp. 263-268.
58. Van der Vleuten and C.P.M., Verwijnen, M., "A system for student assessment," in *Problem-Based Learning: Perspectives from the Maastricht Approach,* (edits. Van der Vleuten, C.P.M and Verwijnen, M.), Thesis-Publisher, Amsterdam, The Netherlands (1990).
59. Posner, G.J., *Analyzing the Curriculum,:* McGraw-Hill, Inc. New York NY (1995).
60. Van Berkel, H.J.M., Nuy, H.J.P. and Geerligs, T., "The influence of progress tests and block tests on study behavior," *Instr. Sci.,* **22,** 317-333(1995).
61. Van der Vleuten, C.P.M., Verwijnen, G.M. and Wijnen, W.H.F.W., "Fifteen years of experience with progress testing in a problem-based learning curriculum," *Med. Teach.,* **18**(2), 103-109(1996).
62. Wasserman, S.I., Kimball, H.R. and Duffy F.D., "Recertification in Internal Medicine: A program of continuous professional development," *Ann. Int. Med.,* **133**(3), 202-8(2000).
63. Edelstein, R.A., Reid, H.M., Usatine, R. and Wilkes, M.S., "A comparative study of measures to evaluate medical students' performance," *Acad. Med.,* **75**(8), 825-833(2000).
64. Miller, G.E., "The assessment of clinical skills/competence/performance," *ibid.,* **65**(9), S63-67(1990).
65. Rethans, J.J., Van Leeuwen, Y., Drop, R., van der Vleuten, C. and Sturmans, F., "Competence and performance: two different concepts in the assessment of quality medical care," *Fam. Pract.,* **7**(3), 168-174(1990).
66. Rethans, J.J., Sturmans, F., Drop, R. and Van der Vleuten, C. and Hobus, P., "Does competence of general practitioners predict their performance? Comparison between examination setting and actual practice," *BMJ,* **303,** 1377-1380(1991).
67. Ram, P., *Comprehensive Assessment of General Practitioners: A Study on Validity, Reliability and Feasibility,* Thesis Publications, Maastricht, Netherlands (1998).
68. Scherpbier, A.J.J.A., van der Vleuten, C.P.M., Rethans, J.J. and van der Steeg, A.F.W., "*Advances in Medical Education*". Kluwer Academic Publishers, Dordrecht, The Netherlands (1997).
69. Newble, D.I., "Assessing clinical competence at the undergraduate level," *Med. Educ.,* **26,** 504-511(1992).
70. Beck, D.E., Boh, L.E. and O'Sullivan, P.S., "Evaluating student performance in the experiential setting with confidence," *Am. J. Pharm .Educ.,* **59,** 236-247(1995).
71. McGuire, C, "Written methods for assessing clinical competence," in *Further Developments in Assessing Clinical Competence,* (edits. Hart, I.R. and Harden, R.M.) Heal Publications, Montreal (1987) pp. 59-75.
72. Van der Vleuten, C.P.M. and Newble, D.I., "How can we test clinical reasoning?" *Lancet,* **345,** 1032-1034(1995).
73. Page, G. and Bordage, G., "The Medical Council of Canada's key features project: A more valid written examination of clinical decision-making skills," *Acad. Med.,* **70**(2), 104-110(1995).
74. Page, G., Bordage, G. and Allen, T., "Developing key-feature problems and examinations to assess clinical decision-making skills," *ibid.,* **70**(3), 194-201(1995).
75. Schuwirth, L., Blackmore, D.E., Mom, E., van den Wildenberg, F., Staffers, J.E.J.H. and van der Vleuten, C.P.M., "How to write short cases for assessing problem-solving skills," *Med. Teach.,* **21**(2), 144-150(1999).
76. Harden, R., Stevenson, M., Downie, W. and Wilson, G., "Assessment of clinical competence using objective structured examinations," *BMJ,* **1**, 447-451(1975).
77. Van der Vleuten, C.P.M. and Swanson, D.B., "Assessment of clinical skills with standardized patients: State of the art," *Teach. Learn. Med.,* **2**(2), 58-76(1990).
78. Van Dalen, J., "Skillslab - A centre for training of skills," in *Problem-Based Learning: Perspectives from the Maastricht Approach,* (edits. Van. der Vleuten , C.P.M.and Verwijnen, M.), Thesis-Publisher, Amsterdam, The Netherlands (1990).
79. Reznick, R.K., Blackmore, D., Cohen, R., Baumber, J, Rothman, A., Smee, S., Chalmers, A., Poldre, P., Birtwhistte, R., Walsh, P., Spady.D. and Berard, J., "An objective structured clinical examination for the licentiate of the Medical Council of Canada: from research to reality," *Acad. Med.,* **68**(10 Suppl), S4-6(1993).
80. Reznick, R.K., Blackmore, D., Dauphinee, W.D., Rothman, A.I. and Smee, S., "Large-scale high-stakes testing with an OSCE: report from the Medical Council of Canada," *ibid.,* **71**(1)(Suppl.), S19-21(1996).
81. Rethans, J.J., Drop, R., Sturmans, F. and van der Vleuten, C, "A method of introducing standardized (simulated) patients into general practice consultations," *Br. J. Gen. Prac.,* **41,** 94-96(1991).
82. Rethans, J.J. Sturmans, F., Drop, R. and van der Vleuten, C, "Assessment of the performance of general practitioners by the use of standardized (simulated) patients," *Br J. Gen. Prac,* **41,** 97-99(1991).
83. Monaghan, M., Vanderbush, R.E., Allen, R.A., Heard, J.K., Cantrell, M. and Randall, J., "Standardized patient use outside of academic medicine: Opportunities for collaboration between medicine and pharmacy," *Teach. Learn. Med,* **10**(3), 178-182(1998).
84. Sibbald, D., "Innovative, problem-based, pharmaceutical care courses for self-medication," *Am. J. Pharm. Educ.,* **62,** 109-119(1998).
85. Austin, Z. and Tabak, D., "Design of a new professional practice laboratory course using standardized patients," *ibid.,* **62,** 271-9(1998).
86. Blake, J.M., Norman, G.R., Keane, D.R., Mueller, B., Cunnington, J. and Didyk, N., "Introducing progress testing in McMaster University's problem-based medical curriculum: psychometric properties and effect on learning," *Acad. Med.,* **71,** 1002-1007(1996).
87. Arnold, L. and Willoughby, T.L., "The quarterly profile examination," *ibid.,* **65**(8), 515-516(1990).
88. Van Berkel, H.J.M., Sprooten, J. and De Graaff, E., "The development of a progress test for a multi-master program health sciences' curriculum," in *Problem-Based Learning as an Educational Strategy,* (edits. Bouhuijs, A.J., Schmidt, H.G. and Van Berkel, H.J.M.), Network Publications, Maastricht, The Netherlands (1993).
89. Shen, L., "Progress testing for postgraduate medical education: A four-year experiment of American College of Osteopathic Surgeons Resident examinations," *Adv. Health. Sci. Educ.,* **5,** 117-129(2000).
90. Patrick, J., *Training, Research and Practice,* Academic Press, Cornwall, UK (1992).
91. Swanson, D.B., Case, S.M. and van der Vleuten, C.P.M., "Strategies for

student assessment," in *The Challenge of Problem-Based Learning,* (edits. Boud, B. and Felleti, G.), Kogan Page Limited, London (1991) pp. 260-273).
92. Swanson, D.B., Case, S.M. and van der Vleuten, C.P.M., "Strategies for student assessment." in *The Challenge of Problem-Based Learning,* (edits. Boud, B. and Felleti, G.), Kogan Page Limited, London (1997) pp. 269-282.
93. Norman, G.R., Tugwell, P., Feightner, J.W., Muzzin, LJ. and Jacoby, J.J., "Knowledge and clinical problem-solving," *Med. Educ.,* **19,** 344-356(1985).
94. Schmidt, H., Norman, G. and Boshuizen, H., "A cognitive perspective on medical expertise: Theory and implications," *Acad. Med.,* **65,** 611-621(1990).
95. Chi, M. T. H., Glaser, R. and Rees, E. "Expertise in problem solving," in *Advances in the Psychology of Human Intelligence, (edit.* Stemberg, R.J.), Lawrence Erlbaum Associates, Hillsdale NJ (1982) pp. 7-76
96. Norcini, J.J., Swanson, D.B., Grosso, LJ. and Webster, G.D., "Reliability, validity and efficiency of multiple-choice question and patient management problem item formats in assessment of clinical competency," *Med. Educ.,* **19,** 238-247(1985).
97. Van der Vleuten, C.P.M., Van Luyk, S.J. and Beckers, H J.M., "A written test as an alternative to performance testing," *ibid.,* **23,** 97-107(1989).
98. Ram, P., van der Vleuten, C, Rethans, J.J., Schouten, B., Hobma, S. and Grol, R., "Assessment of general practice: The predictive value of written-knowledge tests and a multiple-station examination for actual medical performance in daily practice," *ibid.,* **33,** 197-203(1999).
99. Newble, D.I., Baxter, A. and Elsmlie, R.G., "A comparison of multiple-choice tests and free-response tests in examinations of clinical competence," *ibid.,* **13,** 263-268(1979).
100. Norcini, J.J., Swanson, D.B., Webster, G.D. and Grosso, L.J., "A comparison of several methods of scoring patient management problems," Paper presented at the 22nd Annual Conference in Research in Medical Education, Washington DC (1983).
101. Schuwirth, L.W.T., van der Vleuten, C.P.M. and Donkers, H.H.L.M., "Open-ended questions versus multiple-choice questions," in *Approaches to the Assessment of Clinical Competence, Proceedings of the Fifth Ottawa Conference* (Vol. 2, pp. 486-491), (edits. Harden, R., Hart, I.R. and Mulholland, H.), Page Brothers Ltd., Norwich, Great Britain (1992).
102. Schuwirth, L., Van der Vleuten, C.P.M. and Donkers, H.H.L.M., "A closer look at cueing effects of multiple-choice questions," *Med. Educ.,* **30,** 44-49(1996).
103. Schwartz, P.L. and Loten, E.G., "Brief problem-solving questions in medical school examinations: Is it necessary for students to explain their answers?" *ibid.,* **33,** 823-827(1999).
104. Case, S.M. and Swanson, D.B., *Constructing Written Test Questions for the Basic and Clinical Sciences,* National Board of Medical Examiners, Philadelphia PA (1996).
105. Solomon, D.J., Reinhart, M.A., Bridgham, R.G., Munger, B.S. and Starnaman, S., "An assessment of an oral examination format for evaluating clinical competence in emergency medicine," *Acad. Med.,* **65**(9)(Suppl.), S43-44(1990).
106. Turnbull, J., Danoff, D. and Norman, G., "Content specificity and oral certification examinations," *Med. Educ.,* **30,** 56-59(1996).
107. Sibbald, D., "Oral Clinical Skills Examinations: An innovative reinforcing strategy for nonprescription medication courses," *Am. J. Pharm .Educ.,* **62,** 458-463(1998).
108. Roberts, C, Sarangi, S., Southgate, L, Wakeford, R. and Wass, V, "Oral examinations-Equal opportunities, ethnicity and fairness in the MRCGP," *BMJ,* **320,** 370-375(2000).
109. Veloski, J., Rabinowitz, H. and Robeson, M., "Cueing in multiple-choice questions: A reliable, valid and economical solution," in *Research in Medical Education: 1988, Proceedings of the 27th Annual Conference,* AAMC, Washington DC (1988) pp. 195-200).
110. Veloski, J., Rabinowitz, H., and Robeson, M., "A solution to the cueing effects in multiple-choice questions: The Un-Q-format," *Med .Educ.,* **27,** 371-375(1993).
111. Case, S.M. and Swanson, D.B., "Extended-matching items: a practical alternative to free-response questions," *Teach. Learn. Med.,* **5**(2), 107-115(1993).
112. Winslade, N.E., *Assessment of Canadian Forces Physician's Assistants Knowledge of Authorized Pharmaceuticals,* University of Maastricht Press, Maastricht, Netherlands (2000).
113. USMLE. *2001 Examination Bulletin, Step 1, 2 & 3 Content description and sample test materials.* www://usmle.org (2000).
114. Case, S.M. and Swanson, D.B., "Evaluating diagnostic pattern: a psychometric comparison of items with 15, 5 and 2 options," Paper presented at the meeting of the American Educational Research Association, San Francisco CA (1989).
115. Veloski, J.J., Rabinowitz, H.K., Robeson, M.R. and Young, P.R., "Patients don't present with five choices: An alternative to multiple-choice tests in assessing physicians' competence," *Acad. Med.,* **74**(5), 539-546(1999).
116. McGuire, C. and Solomon, C., *Construction and Use of Written Simulation,.* The Psychological Corporation, Chicago IL (1976).
117. Feletti, G.I. and Engel, C., "The modified essay question for testing problem-solving skills," *Med. J. Aust.,* **1,** 79-80(1980).
118. Barrows, H.S. and Tamblyn, R., "The portable patient problem pack (P4). A problem-based learning unit," *J. Med. Educ.,* **52,** 1002-1004(1977).
119. Williams, R., Vu, N, Barrows, H. and Verhulst, S., "Profile of the clinical reasoning test: An objective measure of problem-solving skills and proficiency in using medical knowledge," in *Tutorials in Problem-based Learning,* (edits. Schmidt, H. and De Voider, M.) Van Gorcum Press, Assen, Netherlands (1984).
120. Barrows, H.S. and Pickell, G.C., *Developing Clinical Problem-Solving Skills: A Guide to More Effective Diagnosis and Treatment,* W.W. Norton & Company, Inc., New York NY (1991).
121. Norman, G.R., "Problem-solving skills, solving problems and problem-based learning," *Med. Educ.,* **22,** 279-286(1988).
122. Regehr, G. and Norman, G.R., "Issues in cognitive psychology: Implications for professional education," *Acad. Med.,* **71**(9), 988-1001(1996).
123. Elstein, A.S., Shulman, L.S. and Sprafka, S.A., *Medical Problem-Solving: An Analysis of Clinical Reasoning,* Harvard University Press, Cambridge MA (1978).
124. Swanson, D.B., "A measurement framework for performance based tests," in: *Further Developments in Assessing Clinical Competence,* (edits. Hart, I.R. and Harden, R.M.) Heal-Publications, Montreal (1987).
125. Bordage, G. "An alternative approach to PMP's: The 'key-features' concept," in *Further Developments in Assessing Clinical Competence, Proceedings of the Second Ottawa Conference,* (edits. Hart, I.R. and Harden, R.M.), Can-Heal Publications, Inc., Montreal: (1987) pp.59-75.
126. Bordage, G., Brailovsky, C, Carretier, H. and Page. G.G., "Content validation of key features on a National examination of clinical decision-making skills," *Acad. Med.,* **70**(4), 276-81(1995).
127. Schuwirth, L., van der Vleuten, C.P.M., Mom, E.M.A., van der Waart, T.H.A.M. and Peperkamp, A.G.W., *Direct and Indirect Validity of a Test for Computerized Case-Based Testing [dissertation],* University of Limburg, Maastricht, The Netherlands (1998).
128. NBME and Federation of State Medical Boards of the United States, *USMLE Bulletin of Information,* NBME, Philadelphia PA (2000).
129. Medical Council of Canada, *Objectives for the Qualifying Examination,* Canso Printing Services Inc., Ottawa, Ontario (1999).
130. Charlin, B., Brailovsky, C, Leduc, C. and Blouin, D., "The diagnosis script questionnaire: A new tool to assess a specific dimension of clinical competence," *Adv. Health Sci. Educ.,* **3,** 51-58(1998).
131. Charlin, B., Brailovsky, C, Roy, L., Goulet, F. and Van der Vleuten, C, "The script concordance test: A tool to assess the reflective clinician," *Teach. Learn. Med.,* **12**(4), 189-195(2000).
132. McKinnon, G.E., Pitterle, M.E., Boh, L.E. and Demuth, J.E., "Computer-based patient simulations: Hospital pharmacists' performance and opinions," *Am. J. Hosp. Pharm.,* **49,** 2740-2745(1992).
133. Hepler, CD. and Strand, L.M., "Opportunities and responsibilities in pharmaceutical care," *Am. J. Pharm. Educ.,* **53,** 7S-15S(1989).
134. Hepler, CD., "The Pharmacist in the medication-use process (an introduction to pharmaceutical care)." Proceedings of the Section of Community Pharmacists, World Congress of Pharmacy and Pharmaceutical Science, International Pharmaceutical Federation. September 9-10, 1993. Tokyo, Japan.
135. Perrier D.G, Winslade N.E., Pugsley J.A., Lavack, L and Strand, L., "Designing a pharmaceutical care curriculum," *Am. J. Pharm. Educ.*, **59,** 113-125(1995).
136. Tomechko, M.A., Strand, L.M., Morley, P.C., *et al.* "Questions and answers from the pharmaceutical care project in Minnesota," *Am.* PWm.,**NS35**(4), 30-38(1995).
137. Winslade N.E., Strand, L.M., Pugsley, J.A. and Perrier, D.G., "Practice functions necessary for the delivery of pharmaceutical care," *Pharmacotherapy,* **16**(5), 889-898(1996).
138. Hubal, R.C., Kiazkevich, P.N., Guinn, C.I., Merino, K.D. and West, S.L., "The virtual standardized patient; Simulated patient-practitioner dialog for patient interview training," *Stud. Health. Tech. Inf.,* **70,** 133-138(2000).
139. Issenberg, S.B., McGaghie, W.C., Hart, I.R., Mayer, J.W., Felner, J.M., Petrasa, E.R., Waugh, R.A., Borwn, D.D., Safford, R.R., Gessner, I. H.,

Gordon, D.L. and Ewy, G.A., "Simulation technology for health care professional skills training and assessment," *JAMA,* **282,** 861-866(1999).
140. Tiechen, A., "Competency assessment: Establishing a program," *Clin. Lab. Manage .Rev.,* **13**(5), 275-285(1999).
141. Schwabbauer, M., "But can they do it? Clinical Competency Assessment," *Clin. Lab. Sci.,* **13**(1), 47-52(2000).
142. Ram, P., Van der Vleuten, C.P.M., Rethans, J.J., Grol, R. and Aretz, K., "Assessment of practicing family physicians: Comparison of observation in a multiple-station examination using standardized patients with observation of consultations in daily practice," *Acad. Med.,* **74**(1), 62-69(1999).
143. Ram, P., Grol, R., Rethans, J.J., Schouten, B., Van der Vleuten, C. and Kester, A., "Assessment of general practitioners by video observation of communication and medical performance in daily practice: issues of validity, reliability and feasibility," *Med. Educ.,* **33,** 447-454(1999).
144. Newble, D.I., "Eight years' experience with a structured clinical examination," *ibid.,* **22,** 200-204(1988).
145. Vu, N.V., Barrows, H.S., Marcy, MX., Verhulst, S.J., Colliver, J.A. and Travis, T., "Six years of comprehensive, clinical, performance-based assessment using standardized patients at the Southern Illinois University School of Medicine," *Acad. Med.,* **67,** 42-50(1992).
146. Monaghan, M.S., Gardner, S.F., Hastings, J.K., Reinhardt, G.L., Knoll, K.R., Vanderbush, R.E. and Cantrell, M., "Student attitudes toward the use of standardized patients in a communications course," *Am. J. Pharm. Educ.,* **61,** 131-136(1997).
147. Newble, D. and Jaeger, K., "The effect of assessments and examinations on the learning of medical students," *Med. Educ.,* **17,** 165-171(1983).
148. American Board of Internal Medicine, *Policies and Procedures 2000,* http://www.abim.org/about/p&p.htm (2000).
149. Gerrow, J.D., Chambers, D.W., Henderson, B.J. and Boyd, M.A., "Competencies for a beginning dental practitioner in Canada," *J. Can. Dent. Assoc.,* **64**(2), 94-97(1998).
150. Gessaroli, M.E. and Poliquin, M., "Competency-based certification project. Phase 1: job analysis," *Can. J. Med. Rod. Tech.,* **25**(3), 104-108(1994).
151. American Board of Internal Medicine, *Project Professionalism* and *Strategies for Evaluating Professionalism,* http://www.abim.org/pubs/p2/ (2000).
152. Hammer, D., Mason, H.L., Chalmers, R.K., Popovich, N.G. and Rupp, M.T., "Development and testing of an instrument to assess behavioral professionalism of pharmacy students," *Am. J. Pharm. Educ.,* **64,** 141-151(2000).
153. Swick, H., "Toward a normative definition of medical professionalism," *Acad. Med.,* **75**(6), 612-616(2000).
154. Segall, A., "Generic and specific competence," *Med. Educ.,* 14(Suppl.), 19-22(1980).
155. Friedman, M. and Mennin, S.P., "Rethinking critical issues in performance assessment," *ibid.,* **66**(7), 390-395(1991).
156. Arnold, E., Blank, L., Race, K. and Cipparone, N., "Can professionalism be measured? The development of a scale for use in the medical environment," *ibid.,* **73,** 1119-1121(1998).
157. Boon, H. and Stewart, M., "Patient-physician communication assessment instruments: 1986-1996 in review," *Pat. Educ. Counsel.,* **35**(3), 161-176(1998).
158. Kurtz, S., Silverman, J. and Draper, J., *Teaching and Learning Communication Skills in Medicine,* Radcliffe Medical Press Ltd., Oxon, United Kingdon (1998).
159. Van Dalen, J., Prince, C.J.A.H., Scherpbier, A.J.J.A. and Van der Vleuten, C.P.M., Evaluating communication skills," *Adv. Health. Sci. Educ.,* **3,** 187-195(1998).
160. Van Thiel, J., Van der Vleuten, C.P.M. and Kraan, H., "Assessment of medical interviewing skills: generalizability of scores using successive MAAS-versions," in *Approaches to the Assessment of Clinical Competence,* (edits. Harden, R., Hart, I. and Mulholland, H.), Page Brothers, Norwich, England (1992).
161. Van Thiel, J., van Dalen, J. and Ram, P., "MAAS-Global Manual," Personal communications, J. van Dalen, September, 2000.
162. Researchers in Clinical Skills Assessment, "Consensus statement of the use of standardized patients to evaluate clinical skills," *Acad. Med.,* **68**(6), 475-477(1993).
163. Colliver, J.H., Travis, T.A., Robbs, R.S., Vu, N.V, Marcy, M.L and Barrows, H.S., "Assessment of uniqueness of information provided by post-encounter written scores on standardized patient examinations," *Eval. Health. Prof.,* **15**(4), 465-475(1992).
164. Cohen, R., Rothman, A.I., Bilan, S. and Ross, J., "Analysis of the psychometric properties of eight administrations of an objective structured clinical examination used to assess international medical graduates," *Acad. Med.,* **71**(1 Suppl), S22-24(1996).
165. Clauser, B.E., Ripkey, D., Fletcher, B., King, A., Klass, D. and Orr, N., "A comparison of pass/fail classifications made with scores from the NBME standardized patient examination and Part II examination," *Acad. Med.,* **68**(10 Suppl), S7-9(1993).
166. Newble, D.I. and Swanson, D.B., "Psychometric characteristics of the objective structured clinical examination," *Med. Educ.,* **22,** 325-334(1988).
167. Reznick, R., Smee, S., Rothman, A., Chalmers, A., Swanson, D., Dufresne, L., Lacombe, G., Baumber, J., Polder, P., Levasseur, L., Cohen, R., Mendez, J., Patey, P., Boudreau, D. and Berard, M., "Objective structured clinical examination for the Licentiate: Report of the pilot project of the Medical Council of Canada," *Acad. Med.,* **67,** 487-494(1992).
168. Swanson, D.B. and Norcini, J.J., "Factors influencing reproducibility of tests using standardized patients," *Teach. Learn. Med.,* **1,** 158-166(1989).
169. Cohen, D.S., Colliver, J.A., Robbs, R.S. and Swartz, M.H., "A large-scale study of the reliabilities of checklist scores and ratings of interpersonal and communication skills evaluated on a standardized-patient examination," *Adv. Health Sci. Educ.,* **1,** 209-213(1997).
170. Whelan, G.P., "Educational Commission for Foreign Medical Graduates: lessons learned in a high-stakes, high-volume medical performance examination," *Med. Teach.,* ,**22**(3), 293-296(2000).
171. Regehr, G., Freeman, R., Robb, A., Missiha, N. and Heisey, R., "OSCE performance evaluations made by standardized patients: comparing checklist and global rating scores," *Acad. Med.,* **74**(10)(Suppl.), S135-137(1999).
172. Gray, J.D., "Global rating scales in residency education," *ibid.,* **71**(1 Suppl), S55-63(1996).
173. Cunnington, J.P.W., Neville, A.J. and Norman, G.R., "The risks of thoroughness: Reliability and validity of global ratings and checklists in an OSCE," Adv. *Health Sci. Educ.,* **1,** 227-233(1996).
174. Reznick, R.K., Regehr, G., Yee, G., Rothman, A., Blackmore, D. and Dauphinee, D., "Process rating forms versus task-specific checklists in an OSCE for medical licensure: Medical Council of Canada," *Acad. Med.,* **73**(10 Suppl), S97-99(1998).
175. American Board of Internal Medicine, *Guide to Evaluation of Residents in Internal Medicine,*: ABIM, Philadelphia PA (1999).
176. Bogels, S.M., van Mourik, T.G.C. and Van der Vleuten, C.P.M., "Authentic assessment of interviewing and counseling skills: Effect of testing time per station on generalizability and validity," *Teach. Learn. Med.,* **7**(3), 155-162(1995).
177. Rothman, A.I., Cohen, R. and Bilan, S., "A comparison of short- and long-case stations in a multiple station test of clinical skills," *Acad. Med.,* **71**(1 Suppl), S110-112(1996).
178. Shatzer, J.H,. Wardrop, J.L., Williams, R.G. and Hatch, T.F., "Generalizability of performance on different-station-length standardized patient cases," *Teach. Learn. Med.,* **6,** 54-58(1994).
179. Reznick, R.K., Smee, S., Baumber, J.S., Choen, R., Rothman, A., Blackmore, D. and Berard, M., "Guidelines for estimated the real cost of an objective structured clinical examination," *Acad. Med.,* **68,** 513-517(1993).
180. Wilkinson, T.J., Newble, D.L, Wilson, P.D., Carter, J.M. and Hems, R.M., "Development of a three-centre simultaneous objective structured clinical examination," *Med. Educ.,* **34,** 798-807(2000).
181. Monaghan, M.S., Vanderbush, R.E., Gardner, S.F., Schneider, E.F., Grady, A.R. and McKay A.B., "Standardized patients: an ability-based outcomes assessment for the evaluation of clinical skills in traditional and nontraditional education," *Am. J. Pharm. Educ.,* **61,** 337-344(1997).
182. Weathermon, R.A., Erbele, S. and Mattson, M., "Use of standardized patients as an assessment tool at the end of an ambulatory care rotation," *ibid.,* **64,** 109-113(2000).
183. National Association of Pharmacy Regulatory Authorities, *Competencies Required of Newly-Registered Pharmacists in Canada,* NAPRA, Ottawa, Ontario (1997).
184. Hunt, D.D., "Functional and dysfunctional characteristics of the prevailing model of clinical evaluation systems in North American medical schools," *Acad. Med.,* **67**(4), 254-259(1992).
185. Turnbull, J., Gray, J. and MacFadyen, J., "Improving in-training evaluation programs," *J. Gen. Int. Med.,* **13,** 317-323(1998).
186. Kassebaum, D.G. and Eaglen, R.H., "Shortcomings in the evaluation of students' clinical skills and behaviors in medical school," *Acad. Med.,* **74**(7), 842-849(1999).
187. Brennan, B.G. and Norman, G.R., "Use of encounter cards for evaluation of residents in obstetrics," *ibid.,* **72**(Suppl 10), S43-44(1997).
188. Littlefield, J. and Terrell, C, "Improving the quality of resident performance appraisals," *ibid.,* **72**(10 Suppl) S45-47(1997).
189. Hatala, R. and Norman., G.R., "In-training evaluation during an internal

medicine clerkship," *ibid.,* **74(10),** SI 18-20(1999).
190. Haber, R.J. and Avins, A.L., "Do ratings on the American Board of Internal Medicine Resident Evaluation Form detect differences in clinical competence?" *J. Gen Int. Med.,* **9,** 140-145(1994).
191. Holmboe, E.S. and Hawkins, R.E., "Methods for evaluating the clinical competence of residents in internal medicine: A review," *Ann. Int. Med.,* **129**(1), 42-48(1998).
191. Thompson, W.G., Lipkin, M., Gilbert, D.A., Guzzo, R.A. and Roberson, L., "Evaluating evaluation: Assessment of the American Board of Internal Medicine Resident Evaluation Form." *J. Gen. Int. Med.,* **5,** 214-217(1990).
192. Norcini, J.J. and Day, S. (edits), "Rating forms for the evaluation of individual residents and programs," in *Guide to Evaluation of Residents in Internal Medicine,* American Board of Internal Medicine, Philadelphia PA (1999).
193. *The Mini-CEX: Taking It from the Max,* American Board of Internal Medicine, Philadelphia PA (2000).
194. Field, S., "Passing the MRCGP: Video assessment," *Practitioner.,* **242,** 721-724(1998).
195. Kane, M.T., "The assessment of professional competence," *Eval. Health Prof.,* **15,** 163-182(1992).
196. Hahnemann, B.K., "Journal writing: A key to promoting critical thinking in nursing students," *J. Nurs. Educ.,* **25**(5)**,** 213-215(1986).
197. Wong, F.K.Y., Kember, D., Chung, L.Y.F. and Yan, L., "Assessing the levels of student reflection from reflective journals," *J. Adv. Nurs.,* **22,** 48-57(1995).
198. Challis, M., "AMEE Medical Education Guide No. 11 (revised): Portfolio-based learning and assessment in medical education," *Med. Teach.,* **21**(4), 370-386(1999).
199. Mathers, N.J., Challis, M.C., Howe, A.C. and Field, N.J., "Portfolios in continuing medical education - effective and efficient?" *Med. Educ.,* **33,** 521-530(1999).
200. Pee, B., Woodman, R., Fry, H. and Davenport, E.S., "Practice-based learning: Views on the development of a reflective learning tool," *ibid.,* **34,** 754-761(2000).
201. Dennick, R., "Case Study 2: Use of logbooks," *ibid.,* **34**(Suppl), 66-68(2000).
202. Snadden, D., "Portfolios - Attempting to measure the unmeasurable?" *ibid.,* **33,** 478-479(1999).
204. Schon, D.A., *Educating the Reflective Practitioner: Towards a New ign for Teaching and Learning in the Professions,* Jossey, Bass, San Francisco CA (1987).
205. Kolb, D.A., *Experiential Learning: Experience as the Source of Learning and Development,* Prentis Hall, englewood Cliffs NJ (1984).
206. Burton, A.J., "Reflection: Nursing's practice and education panacea," *J. Adv. Nurs.,* **31**(5)**,** 1009-1017(2000).
207. Pitts, J., Coles, D. and thomas, P., "Educational portfolios in the assessment of general practice trainers: Reliability of assessors," *Med. Educ.,* **33,** 515-520(1999).
208. Fabb, W., *The Examination and Assessment System of the RACGP. A Manual for Examiners,* College of General Practitioners, Australia (1991).

APPENDIX. MCMASTER ENCOUNTER CARD FROM HATALA AND NORMAN, 1999(189)

(Reprinted with permission from Academic Medicine and the Association of American Medical Colleges)

Date(yy/mm/dd): Student Name:_____ ID #:_____
Evaluator Name: _____ Evaluator's Signature:_____
Evaluator (circle one): Attending PGY-2 PGPY-3 PGY-4
Hospital (circle one): (1) MUMC (2) HGH (3) Henderson (4) St. Joseph's

TO BE COMPLETED BY EVALUATOR Principle Focus of Encounter (check one):

		Yes	No
(1)	Clinical Skills: History Directly observed?	( )	( )
(2)	Clinical Skills: Physical Directly observed?	( )	( )
(3)	Professional Behavior Directly observed?	( )	( )
(4)	Case Presentation: (circle one)	written	verbal
(5)	Problem Formulation: Diagnosis		
(6)__	Problem Formulation: Therapy		
(7)__	Other(describe): _____		

Rating of Encounter

[ ]	[ ]	[ ]	[ ]	[ ]
Unsatis-factory	Performed BELOW level of average clinical clerk	Performed AT level of Average Clinical Clerk	Performed ABOVE level of average clinical clerk	Performed at level of INTERN average
(1)	(2)	(3)	(4)	(5)

Comments on Student Performance: